I T H E A

International Journal INFORMATION THEORIES & APPLICATIONS

2020 Volume 27 Number 1

International Journal

INFORMATION THEORIES & APPLICATIONS
Volume 27 / 2020, Number 1

International Journal "INFORMATION THEORIES & APPLICATIONS" (IJ ITA)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITA welcomes scientific papers connected with any information theory or its application. IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA are given on *www.ithea.org*.

Responsibility for papers *published in* IJ ITA belongs to authors.

International Journal "INFORMATION THEORIES & APPLICATIONS"
Vol. 27, Number 1, 2020

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with: University of Telecommunications and Posts, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain,
Hasselt University, Belgium, University of Perugia, Italy,
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia
St. Petersburg Institute of Informatics, RAS, Russia,

Printed in Bulgaria
Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: office@ithea.org
Technical editor: **Ina Markova**

ISSN 1310-0513 (printed) **ISSN 1313-0463 (online)**

USING VISUAL ANALYTICS AND K-MEANS CLUSTERING FOR MONETISING LOGISTICS DATA, A CASE STUDY WITH MULTIPLE E-COMMERCE COMPANIES

Hamzah Qabbaah, George Sammour, Koen Vanhoof

Abstract:

Logistics companies possess and collect a large amount of data on the shipments they perform while at the same time facing a challenge to understand their complicated market better. Therefore, investigating whether large databases gathered by logistics companies on their e-commerce partners could be monetised as a business service and how this could eventually be achieved is an important research venture. In this paper we used visual analytics and k-means clustering to see whether the data could be structured and presented in a monetisable way, while at the same time adhering to the quality characteristics necessary for doing so: reliable, accurate, relevant, segmented, secured and anonymized. Results show that is clearly the case for the database we investigated and contained 85989 transactions. Using a semi-structured interview with several key managers of both the logistics company and its e-commerce partners, a business-model canvass was developed that indicates the necessary elements for this venture and the right mindset to manage the process. We can confidently conclude that all elements are present to answer the monetisability question positively and to pretend that given the right visualization and confidence between the companies the process could very well be profitable.

Keywords: *Data monetisation, Visual analytics, K-means clustering, Logistics, E-commerce.*

1. Introduction

The concept of data monetisation is to be situated in the broader context of data sharing between companies Business to Business (B2B) Data Sharing. The exchange of data between organisations producing and holding data constitutes the 'data market' (data supplier companies, representing the supply side of the market) and organisations using or re-using data (data user companies, representing the demand side of the market) is situated at the core of B2B data sharing, representing the exchange of data resources with multiple applications or users in different organisations (Carnelley 2018).

Many companies use the data produced in their value chain only internally to improve productivity, manage costs, or improve customer relationships. They do not realise that data are also an asset, when they are used to create new revenue externally. Data monetisation is all about this way of exploiting data (Carnelley 2018).

Findling et al. (Serge Findling 2018) identified three primary paths towards realising that value for data: (1) data sale or licensing creating immediate revenue; (2) bundling data with other services or products thus creating extra value for these services; and (3) exchanging premiums/trade advantages or discounts for data. Our research only deals with the first two ways of monetising data.

2. Background

Several definitions of data monetisation have been put forward in literature. Moore (Moore 2015) stresses that through data monetisation data are transformed into a source of profit. Najjar and Kettinger (Najjar and Kettinger 2014) describe data monetisation more clearly by stating that "Data monetisation happens when the intangible value of data is converted into real value, usually by selling them". The definition by Fred (FRED 2017) can be considered as being more generic, since in his eyes the revenue will not only be obtained by selling the data but also by creating different applications and services. His definition identifies data monetisation as: "The revenue generation

with and out of data and data-derived information-based products and services". Finally, Thomas and Leiponen (Thomas and Leiponen 2016) have developed a categorisation scheme for monetising data. They identify three categories : (1) Selling data; (2) Providing insights or analyses and (3) Creating new services.

Scientific literature indicates a number of factors favouring the creation of value through data sharing (Liu, Ren et al. 2009, Najjar and Kettinger 2014, Bonneau 2015, Bataineh, Mizouni et al. 2016). They can be subdivided in three areas: technological factors, business factors and execution factors.

All three areas may have relevance either to the quality of the data that can be shared and monetised or to the mind-set and processes going on in the company.

Companies have to assess the value of their data carefully. From a theoretical point of view, four characteristics or considerations are relevant (platform 2018) (Bataineh, Mizouni et al. 2016).

The data have to be: (1) Reliable and accurate, (2) Relevant, (3) Segmented and (4) Secure and anonymised.

Data monetisation creates a new business model for the company involved. Data are not only used to run the business anymore, they become a product/service that can be sold to partners and generates an income larger than the cost of creating and gathering the data (Liu, Ren et al. 2009, Najjar and Kettinger 2014). This signifies that not only a correct way of connecting and sharing the data with the partners has to be developed, but also a route to sell the data in the best possible way. The connecting and sharing of the data is a technological factor in terms of the conditions for success, whereas the selling mode is both a business and execution factor (Najjar and Kettinger 2014).

Connecting and sharing information with suppliers necessitates an improved technical capability by developing a supplier portal (hosted by a third-party analytics firm) that allows to share information with partners (Najjar and Kettinger 2014, Bonneau 2015).

3. Literature Review

Gottlieb and Rifai (Company 2017) have indicated the rapid development of the data sharing and monetisation market. An enquiry among 530 executives and senior level managers revealed that data monetisation was widely seen as a new means of generating revenue. 41% of the respondents had begun doing so in recently. Everis (Catarina Arnaut 2018) indicated that among companies starting to share their data with other companies a majority of 52% asked some form of remuneration for this sharing, while 40% were willing to share their data to a limited number of partners for free. 8% shared data freely to a wider audience. The generated income (Catarina Arnaut 2018) is very diverse however and ranges from 5000 euros (for one third of all respondents) to more than one million euros (in one fifth of these cases), depending also on the size of the companies in the sample. These investigations prove that data monetisation is slowly becoming part of the business world.

Literature mentions a number of examples. The most explicit one is Dawex (Carnelley 2018). Founded in 2015, this French data marketplace operates a website by connecting companies selling or buying data, using standard enterprise software. Dawex has managed in two years to progressively enlarge its data offer to a wide array of industries, from automotive, to energy, from agriculture to retail, healthcare and, more recently, financial services. 2,000 companies are connected to the platform. 45% of them are situated in Europe, 38% in the United States, and 17% in Asia.

All other studies are situated in specific sectors of industry. Perrons and Jenssen examined existing data management practices in the upstream oil and gas industry and compare them to practices and philosophies that have emerged in organisations that are leading the way in Big Data. The comparison showed that this kind of data can be regarded as a valuable asset, although they are frequently just regarded as descriptive information about a physical asset (Perrons and Jensen 2015). Bataineha et al. (Bataineh, Mizouni et al. 2016) investigated the use of data gathered from customers in the mobile phone market. 'Mobile phone-based sensing' is a new business practice aiming at using smart phones to answer sensing requests and collect useful data. A

wide variety areas ranging from health-care applications (think actually of the contact tracing apps developed in covid-times) to pollution monitoring are benefiting from these data. Xu et al. (Xu, Qiu et al. 2019) have studied the use of data-driven logistics in commerce from the perspective of risk management. The paper focuses on quantitative operational risks in E-commerce. These operational risks mainly refer to the risks owing to supply/ demand uncertainties, human mistakes and accidents that would decrease the service level or threat the normal operations.

To the best of our knowledge, monetisation of personal data has not been studied extensively, except from the angle of privacy concerns (Laudon 1996, Bélanger and Crossler 2011, Li and Raghunathan 2014) with a focus on organisations as data owners. Authors in (Li and Raghunathan 2014) adopted an economics-based approach which addresses the issue of disseminating sensitive data to a third party data user (Bataineh, Mizouni et al. 2016).

4. Problem statement and Research question

Logistics companies possess large amounts of data because they face the challenge to understand their complicated market better (Qabbaah, Sammour et al. 2019). This research focuses on the possibility of monetising the large databases by logistics companies gathered on the shipments of their partner companies that are mostly e-commerce companies. The economical relevance of such an approach is that this may allow e-commerce companies to allocate some of their limited resources to manage their data more effectively.

Knowing how logistics companies that possess such huge data bases can effectively offer the service of improving the knowledge of their business partners with respect to their business and customers is an interesting field of research. In other words, can they really market their data to these partners? For this purpose the data have to be instrumental in nature (giving customers a better idea about some business questions). Visual analytics using statistics and charts can certainly be used to describe the data and to understand the

market of logistics companies better. The lack of previous research in this field makes this effort very important.

The contribution of this paper is consequently to investigate whether value can be derived in an international context from extensively looking into the monetisation possibility of specific logistics data of e-commerce companies (a field and combination that has not been studied before). The following research question will be addressed **'What is the possibility of data to be monetised in the logistics sector?'**

5. Methodology

In order to answer our research question, we start from a real life dataset. Different methodological strategies were applied. First we will develop in depth statistics and visualization charts on the data aimed at showing in which way they could help in getting a clearer understanding about the dataset. We will use visual analytics in doing so. The most important example of this is e-companies' market analyses. The purpose is to see whether this could enable logistics partner companies to understand their competitiveness in more depth and help them in improving their image and sales volume by looking at creative ways to develop new unique selling propositions.

Second, we will use K-means clustering on the data for segmentation of transactions purposes. This is useful for the e-commerce companies as well as for the customers since the clustering results show the prevalent logistics variable combinations.

Third, we will create a semi-structured interview to be held with the senior decision makers of the partner companies and of the logistics company involved to get their opinion about the data monetisation concept. We will confront the interviewees with some results to find out their reaction and eventually some evidence of the monetisability of our data.

In the next paragraphs we will explain these methodologies.

Data visualization involves "Presenting data in graphical or pictorial form which makes the information easy to understand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information" (Sadiku, Shadare et al. 2016). Whereas, visual analytics is defined as "The science of analytical reasoning facilitated by visual representations used within a personal context" (Huang, Tory et al. 2015). Therefore, visual analytics goes one step further than data visualization. It can be considered as an integral approach combining visualization and data analysis. Visual analytics integrates data visualization methodology with information analytics, geospatial analytics, and scientific analytics. Therefore, visual analytics benefits from methodologies developed in the fields of statistical analytics, data management, knowledge representation, and knowledge discovery (Pak Chung Wong 2004). It is not likely to become a separate field of study, but its influence will spread over the research areas it comprises

K-means clustering is used often for segmentation purposes. This method offers a real-time solution for the development of distributed interactive decision supports since it allows the consumer to model his/her preferences along multiple dimensions, such as product information and logistics route and then produces data clusters of the products-logistics combinations retrieved to enhance marketing decisions (Papamichail and Papamichail 2007).

The main objective of this algorithm is to partition the dataset into k clusters in which each instance belongs to the cluster with the nearest mean. It is suitable for large datasets and offers ease of implementation and high speed performance (Carmona, Ramírez-Gallego et al. 2012). The K-Means algorithm starts from k central point's chosen randomly. Every instance is assigned to the closest central point. Next, the heuristic performs a reassignment of the central points. The algorithm finally stop when the assignments of the individual instances no longer change (Kotu and Deshpande 2019). The algorithm follows five steps.

The first step initiates k random centroids. The number of clusters k should be specified by the user.

Step 2 consists in assigning data points. Once centroids have been initiated, all the data points are assigned to the nearest centroid to form a cluster. In this context the 'Nearest' is calculated by a proximity measure. The Euclidean distance measurement is the most common proximity measure used in this respect. The Euclidean distance between two data points X (x1, x2,…xn) and C (c1, c2,…cn) with n attributes is given by:

$$\text{Distance d} = \sqrt{(\text{x1} - \text{c1})^2 + (\text{x2} - \text{c2})^2 + \cdots + (\text{xn} - \text{cn})^2} \qquad (1)$$

In a third step new centroids are calculated. For each cluster, a new centroid is calculated, which is also the prototype of each cluster group. This new centroid is the most representative data point of all data points in the cluster. Mathematically, this step can be expressed as minimizing the sum of squared errors (SSE) of all data points in a cluster to the centroid of the cluster. The overall objective of the step is to minimize the sum of squared errors of individual clusters. The SSE of a cluster can be calculated by the following equation:

$$\text{SSE} = \sum_{\text{i=1}}^{\text{k}} \sum_{\text{xj} \in \text{ci}}^{\text{k}} ||\text{xj} - \mu\text{i}||^2 \qquad (2)$$

where Ci is the ith cluster, j are the data points in a given cluster, μi is the centroid for ith cluster, and xj is a specific data object. The centroid with minimal SSE for the given cluster i is the new mean of the cluster. The mean of the cluster can be calculated by:

$$\mu\text{i} = \frac{1}{\text{ji}} \sum_{\text{x} \in \text{ci}}^{\text{k}} \text{X} \qquad (3)$$

where X is the data object vector (x1, x2,… xn).

Step 4 is a repeated assignment and calculation of new centroids. Once the new centroids have been identified, assigning data points to the nearest centroid is repeated until all the data points are reassigned to new centroids.

Step 3 and step 4 are iterative until no further change in assignment of data points happens or, in other words, no significant change in centroids is noted anymore. The final centroids are declared the prototype data objects or vectors and they are used to describe the whole clustering model. Each data point in the dataset is now tied with a new clustering ID attribute that identifies the cluster.

Since monetising data is relatively new so that data on the revenue generated and the marketing for it do not exist readily in the logistics sector, we finally wanted to investigate the potential for monetisation. We used a semi-structured interview technique to develop some insight into this point. We opted for this research methodology as it can provide qualitative data, which in this case is more relevant than quantitative ones.

Our enquiry was subdivided in two parts, one for e-companies and one for the logistics provider (which possesses the database). In our interviews we spoke with the manager responsible for the transactions and relationship management with the logistics company. The e-companies were selected by the logistics company on the basis of some criteria: being situated in Jordan, important as a partner and using the logistics services for a number of different items or product categories. These criteria were selected because they offered a variety of different situations in which these companies are operating. For reasons of privacy, their names and activity fields are not disclosed. Our questions were based on previous research (Laitila 2017, Derwisch 2019). They reflect the reality of the data used by the e-companies and the use they can make of it.

We have to stress that this selection is not without danger. The relationship with the logistics company of the managers of the e-companies and the involvement of the logistics company in selecting the interviewees may influence the results in a positive way. Nevertheless, we believe that this bias is limited as we did not mention to the interviewed managers of the e-companies that they were selected because of their relationship with the logistics company and the

questions were phrased in a more general way than to mention logistics databases.

6. Data

Logistics companies that ship products sold online have a huge amount of detailed market basket data available. They contain information on sets of items that buyers acquire, whether the items are bought together or at different times, the physical characteristics of the products they buy, the type of products they like to buy, the suppliers they like to buy from, the payment mode they use, the logistics route used (country of origin and country of destination), the brands they select and price levels that trigger their buying and so on. Moreover, the online buying behaviour of the customers is also registered in the data base. They are however rarely taken into account in a combined product/ customer categorisation effort.

The data used in this research were obtained from a logistics services company situated in the Middle East. Cleaning, merging tables and pre-processing of the data were applied in order to obtain the final data set. The total number of transactions in the final dataset is equal to the size of the sample (n=85959). Table 1 below shows the variables, the type of data they represent and the description of each of the variables used in our research.

Table 1. Data Descriptions

Variable	**Data type**	**Data Description**
Variables in the original dataset		
ID	Integer	The ID of the order
Weight	Double	The weight of shipment

CODValueUSD	Double	The amount of cash on delivery
Payment	String	Type of payment. P: prepaid, C:cash, 3:third party, F:free
Destination	String	Destination of shipment
Origin Country	String	Country of origin of the shipment
DestCountry	String	Country of destination of the shipment
ShipperID	Integer	The ID of the E-commerce companies
CODFlag	Boolean	Cash on delivery flag
'Consignee Tel'	Integer	The telephone number of the customers
Added variables		
Weight In KG	Double	Total weight in KG
Value USD	Double	The price of the goods in the shipment in US Dollar
Product Group name	String	Product group name of the shipment
Product Group ID	Integer	Product group ID
Customs class	Boolean	If the shipment subject to customs or not

Tableau Software' was used to visualize the data. We wanted to find the distributions of transactions in our data on the bases of cities and countries of shipments, how the product categories and returned products distributed on the basis on countries, customers, e-commerce companies and customs class.

Therefore, visualizing the different attributes dimensions, such as location, products, customs class, customers and e-commerce companies was necessary to prepare the data in a format that might be monetisable.

7. Visual analytics results

Table 2 represents the summary of the data we will visualize in this section. It indicates the number of the transactions according to the dimensions (variables) mentioned.

Table 2: The visualization of the different dimensions in this section.

Number of transactions according to	Dimensions (variables)
Destination countries, origin countries and destination cities.	Country (origin, destination), Destination cities.
Products transferred to the country of destination	Products←→ Destination countries
Products transferred to the city of destination	Products ←→ Destination cities
Products transferred from country of origin	Products←→ Origin countries
E-commerce companies have orders transferred to destination countries	e-commerce companies←→ Destination countries
Customers have orders transferred to	Customers←→ Destination

destination countries	countries
Product categories shipped by the customers	Customers ←→ Products
Product categories transferred to destination countries from the countries of origin.	Origin countries→ Destination countries→ Products
Retuned orders by the country of destination	Return products ←→ Destination countries
Retuned orders by the city of destination	Return products ←→ Destination cities
Retuned orders by the e-commerce companies	Return products ←→ E-commerce companies
Retuned orders by the customers	Return products←→ Customers
Customs class by country of origin	Origin countries → Customs
Customs class by country of destination	Destination countries → Customs
Customs class by Product categories	Products→ Customs
Customs class by product categories transferred from country of origin to destination country	Destination countries→ Origin countries→Products→Customs

We will show few examples from Table 2 in the following Figures.

Figure 1 presents the top 10 products transferred to all destination countries together. We can see that 'Apparel' has the highest percentage with 32%, followed by 'Beauty supplies' and 'watches' with 7% and 5%, respectively. Companies can learn from this figure what are the most products ordered by the customers in the our destination countries.

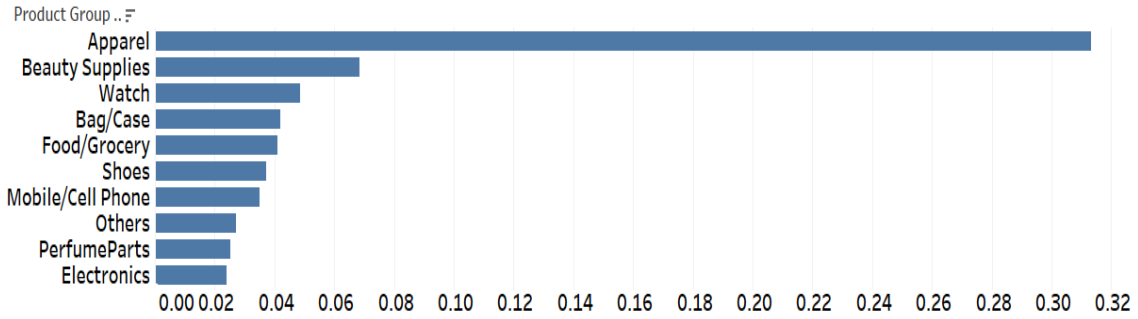

Figure 1: Top 10 products shipped.

Figure 2 presents the percentages of the transactions according to the product categories transferred to Jordan for the top four countries of origin. 'Apparel' is mostly transferred from the 'US' to Jordan (29%), 'Beauty supplies' with 9% is also transferred most frequently from 'US', while shipments from 'GB' are mostly 'Apparel' and 'Home supplies' with 4% and 2%, respectively. Companies can learn from this figure that the customers in Jordan prefer to order their products such as 'Apparel' and ' Beauty supplies' from 'US'.

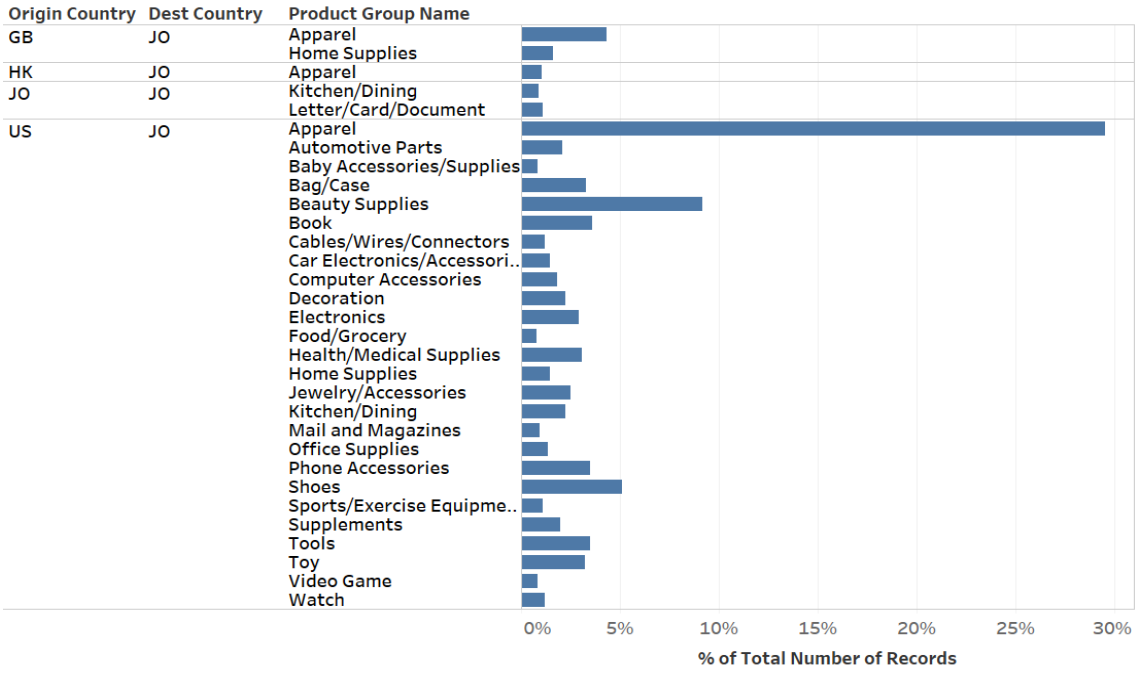

Figure 2: Product categories shipped to Jordan according to the top four countries of origin (expressed in percentages).

Figure 3 shows the percentages of the transactions according to the returned orders by the country of destination and city of destination. 'Jeddah' and 'Riyadh' in Saudi Arabia represent 35% and 34%, respectively, while 'Dubai' represents more returned orders than 'Abu Dhabi' in the 'UAE'.

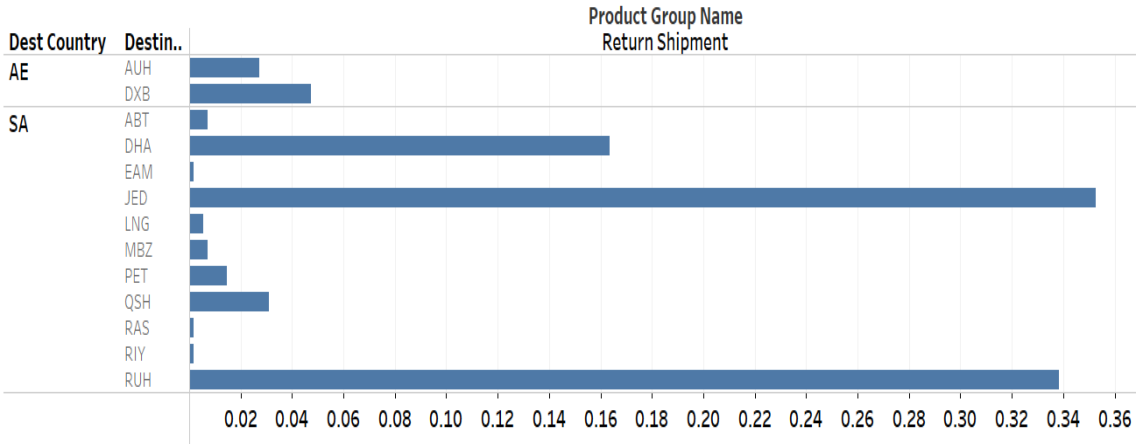

Figure 3: The returned orders according to the city of destination.

Customs variable has been transformed using the 'Customs Value' variable. The variable contains two classes: 'Yes': if the order is subject to customs tariffs and 'No': if the order is not subject to customs tariffs.

Figure 4 shows the percentages of the transactions according to the customs (Class 'Yes') for the most frequently send product categories to Jordan for several countries of origin. Companies can see that 'Apparel' and 'Beauty supplies' transferred from the 'US' to Jordan have the highest percentage of customs application class with 27% and 9%, respectively, while 'Apparel' has the highest percentage of customs transferred from the 'UK' with 4%.

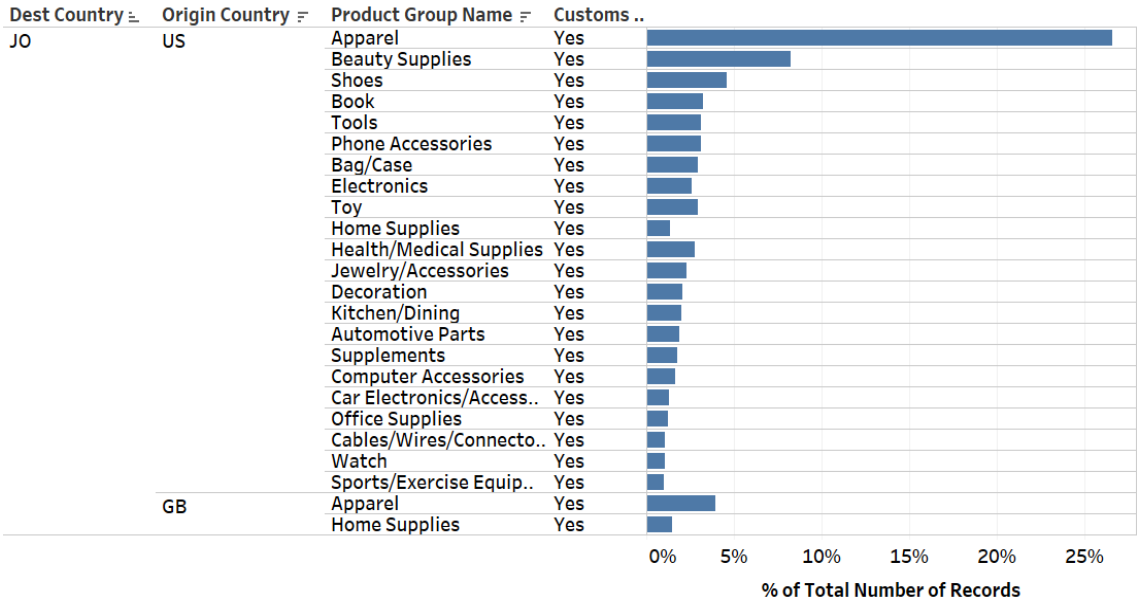

Figure 4: Customs (class 'Yes') according to the most product categories transferred to Jordan (expressed in percentages).

8. E-commerce companies market analyses and results

The visual analytics of our data can fully be understood when the weight of the different variables (products, countries and customers) relative to the total number of transactions is clearly mentioned. This is much like saying that the importance of a product in a certain market can be grasped when compared to

the total market of competitive alternatives. This concept must be visualized as well. Therefore, we have tried to explain the relative weight of the different products-countries-customers in as much detail as possible by clearly describing what have been visualized.

The values were calculated as follows. First, the time period of all the data in the dataset was the same for the companies, products and customers, which makes our comparison accurate on the time level. Second, as an example, we calculate for instance the company's total transactions for the different types of products. Third, we divide the number of transactions of each company by the total number of the product transactions. We also apply this procedure to calculate the countries market for the different products categories.

Some examples are shown below.

Figure 5 presents the destination countries market analysis according to the eight most common products transferred. Section 5A shows that Saudi Arabia has the biggest market for 'Apparel' product with 90.34% following by Jordan and UAE with 6.79% and 2.87%, respectively. Section 5G shows that UAE has bigger market than Jordan for Product 'Letter/card/Document' with 9.57% and 5.54%, respectively, but still Saudi Arabia has the highest percentages with 84.9%. and so on.

Figure 6 presents the e-commerce companies relative importance on the market on the basis of the products transactions for one of the destination countries, namely Jordan.

E-company '15037' has the highest market importance for 'Apparel', 'Bag/Case', 'Beauty supplies', 'Book', 'Food/Grocery', 'Jewellery Accessories' and 'shoes' with 69%, 86%, 92%, 82%, 77%, 85% and 79%, respectively. Whereas e-company '197483' has the highest market importance for 'letter/ card/ document' product with 40%. Companies can learn from this figure how the results of the market importance of the products are distributed between the top five e-companies in Jordan.

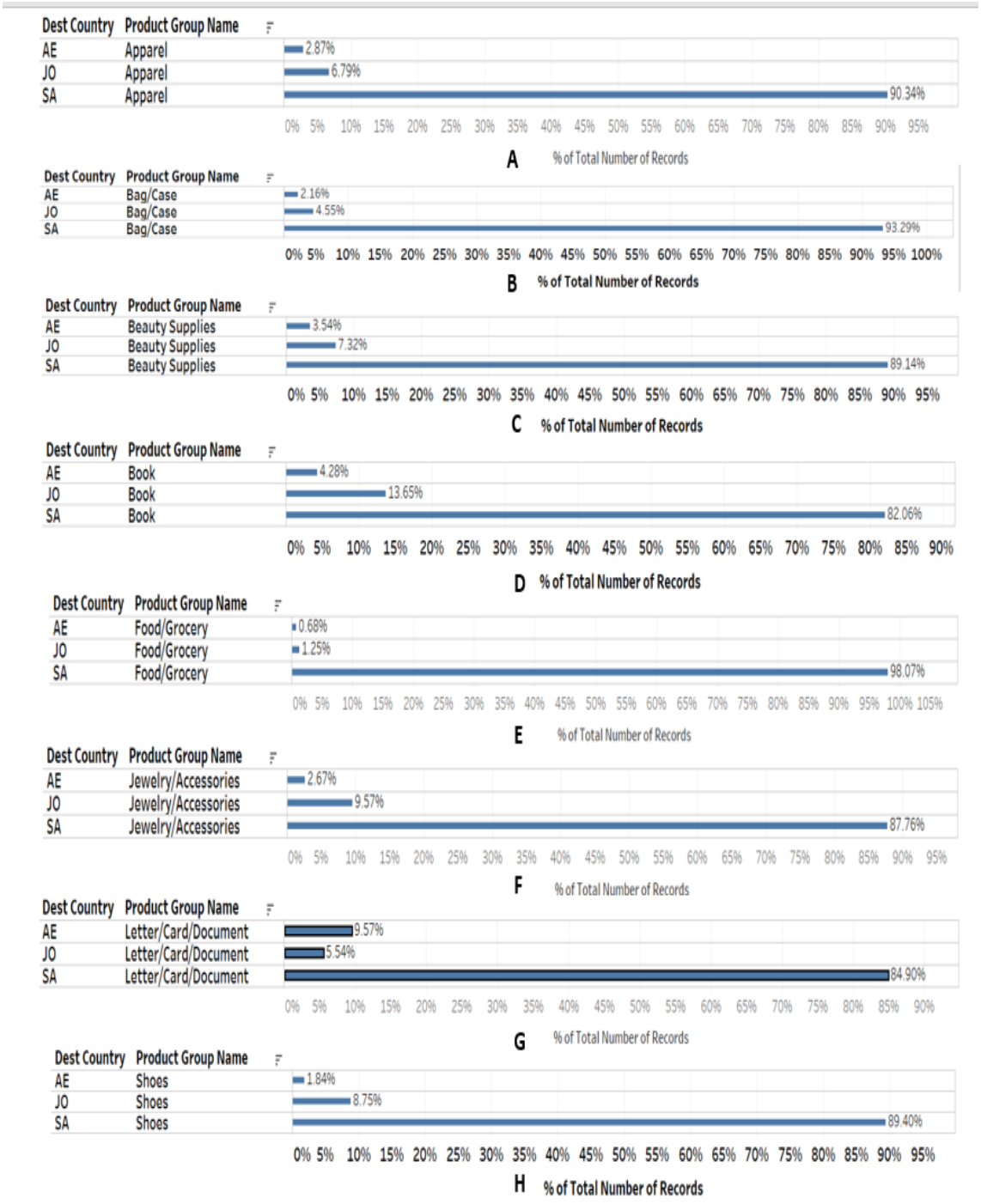

Figure 5: Countries of destination market according of the eight products most commonly ordered (expressed in percentages).

These results could provide e-companies with a clear vision on their position in the market according to products and locations. E-companies can identify and

know their competitors in the market in general or on the basis of some products particularly.

A

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Apparel	69.34%
	64450	Apparel	16.50%
	565631	Apparel	7.92%
	546181	Apparel	1.42%
	565363	Apparel	1.20%

B

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Bag/Case	85.98%
	565631	Bag/Case	7.32%
	64450	Bag/Case	4.27%
	503568	Bag/Case	1.22%
	391989	Bag/Case	0.61%

C

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Beauty Supplies	92.09%
	64450	Beauty Supplies	5.58%
	502065	Beauty Supplies	0.70%
	580389	Beauty Supplies	0.47%
	599148	Beauty Supplies	0.47%

D

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Book	81.91%
	64450	Book	16.49%
	366180	Book	0.53%
	502065	Book	0.53%
	548247	Book	0.53%

E

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Food/Grocery	77.27%
	64450	Food/Grocery	20.45%
	586537	Food/Grocery	2.27%

F

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Jewelry/Accessories	84.50%
	64450	Jewelry/Accessories	12.40%
	601479	Jewelry/Accessories	1.55%
	557424	Jewelry/Accessories	0.78%
	565363	Jewelry/Accessories	0.78%

G

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	197483	Letter/Card/Document	39.77%
	362662	Letter/Card/Document	13.64%
	64450	Letter/Card/Document	11.36%
	15037	Letter/Card/Document	7.95%
	594857	Letter/Card/Document	6.82%

H

Dest Country	Shipper ID	Product Group Name	% of Total Number of Records
JO	15037	Shoes	78.93%
	64450	Shoes	14.29%
	565631	Shoes	2.50%
	565363	Shoes	2.14%
	585583	Shoes	1.07%

Figure 6: E-commerce companies market importance on the basis of the products transactions for Jordan (expressed in percentages).

9. Clustering Results

The clustering model focuses on clustering the logistics transactions on the basis of logistics variables such as product type purchased, customs classes, weight of the shipment, value of goods bought and location of shipments from origin countries to destination cities. This segmentation shows the prevalent logistics variable combinations. These data have shown to be of particular interest to the e-commerce partners and could be monetisable since an improved focus on customers interested in these combinations is possible. They can be served better.

In order to find the best cluster fit experiments we analysed the data for 2 to 5 clusters. We used ANOVA and the Calinski-Harabasz criterion to assess the clustering quality. Calinski-Harabasz has the option that if a user does not specify the number of clusters, the number of clusters will be picked corresponding to the first local maximum of the Calinski-Harabasz index automatically. Then the best fit in each of these models was selected by comparing the clusters results from k=2 till k=5 taking into consideration the evaluation results.

The variables used in our model are 'Value USD', 'Weight In KG', 'Product Group Name', 'Customs', 'Country of origin', 'Country of Destination', 'Destination' and 'Payment type'. The data include all three countries of destination (Jordan, UAE and Saudi Arabia) (N=85,959).

Table 3 shows the results of the 5-cluster solution. The customs class in clusters-1,2 and 4 has been identified as 'Subject to customs' while the customs class in cluster-3 and 5 is 'Not subject to customs'. Both clusters-1 and 2 have the 'Apparel' product as the most frequently shipped product to 'Riyadh' in Saudi Arabia, while the most frequently transferred products in clusters 3, 4 and 5 are 'Food/Grocery', 'Mobile/Cell Phone' and 'Letter/Card/Document'. These transactions are shipped respectively to 'Jeddah' and 'Amman'. The most common transactions in cluster-3 and 5 with a customs class 'Not subject to customs' are shipped within Saudi Arabia and Jordan. Cluster-3 has the highest average of the weight with 9.1 KG, and cluster-2 has the highest average of the price of the shipment with 156 dollar.

Table 3: The results of the 5-cluster solution.

	Attributes/ Clusters	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**
	Number of Items	39882	25073	7731	13010	262
	Avg. Weight In KG	1.3013	3.8994	9.0847	1.6742	1.2353
	Avg. Value USD	75.161	155.5	107.98	110.49	37.12
	Customs	1	1	0	1	0
Most Common	Product Group Name	Apparel	Apparel	Food/ Grocery	Mobile/Cell Phone	Letter/Card/ Document
	Country of Destination	SA	SA	SA	SA	JO
	Destination	RUH	RUH	JED	JED	AMM
	Payment	C	P	P	P	F
	Country of origin	US	HK	SA	SA	JO

Notes: RUH: Riyadh, JED: Jeddah, AMM: Amman. SA: Saudi Arabia, US: USA, HK: Hong Kong, JO: Jordan. Customs class: 1- Subject to customs. 0- Not subject to customs.

Payment: C: cash, P:prepaid, F:Free

Table 4 shows the results of the analysis of the variance (ANOVA) test for our clustering Model.

Table 4: The results of the analysis of the variance test

Number of clusters	Variable	F-statistic	P-value
2-clusters	Avg. Weight In KG	1689	0.000
	Avg. Value USD	698.3	0.000
3-clusters	Avg. Weight In KG	601.9	0.000
	Avg. Value USD	445.4	0.000
4-clusters	Avg. Weight In KG	1503	0.000
	Avg. Value USD	423	0.000
5-clusters	Avg. Weight In KG	1169	0.000
	Avg. Value USD	321.3	0.000

The results of the analysis of variance (ANOVA) of the different cluster solutions show that the p-value is <0.001 for the continuous variable 'Avg.Total Value USD' and 'Avg. Weight In KG'. The results of the Calinski-Harabasz test indicate that k=5 is the best cluster fit. We can see that the number of items in cluster-5 is lower than in the other four clusters. Cluster-5 is the only cluster showing 'Jordan' as a country of destination while the other four clusters have 'Saudi Arabia' as a country of destination. Moreover, in cluster-5, the most frequently product shipped is 'Letter/Card/Document', which is different from the other clusters.

The model has two clusters with a customs class 'Not subject to customs' and three clusters with a customs class 'Subject to customs'. The model has clusters with three different product groups, three different destination cites, three different payment methods and four different country of origins.

The following two figures show the percentages of the highest number of transactions of the 5-cluster solution in detail, four variables are shown in these Figures. The x-axis contains the total number of transactions, whereas the y-axis contains the following shipment variables: customs class, product categories, country of origin and destination city, while the percentages of the total number of transactions are presented per y-axis variables for the cluster group. We will show the results of two clusters only. Cluster-1 with the highest number of orders subject to customs and cluster-3 with the highest number of orders not subject to customs. From Figure 7 companies can learn that in cluster-1 the highest percentages of the transactions were shipped from 'US' to 'RUH', 'JED' and 'AMM', respectively for the category 'Apparel' with a custom class 'Subject to customs'.

Companies can learn from Figure 8 that in cluster-3 the highest percentages of the transactions are transferred within 'SA' for the category 'Food/Grocery', and the customs application class was 'Not subject to customs'.

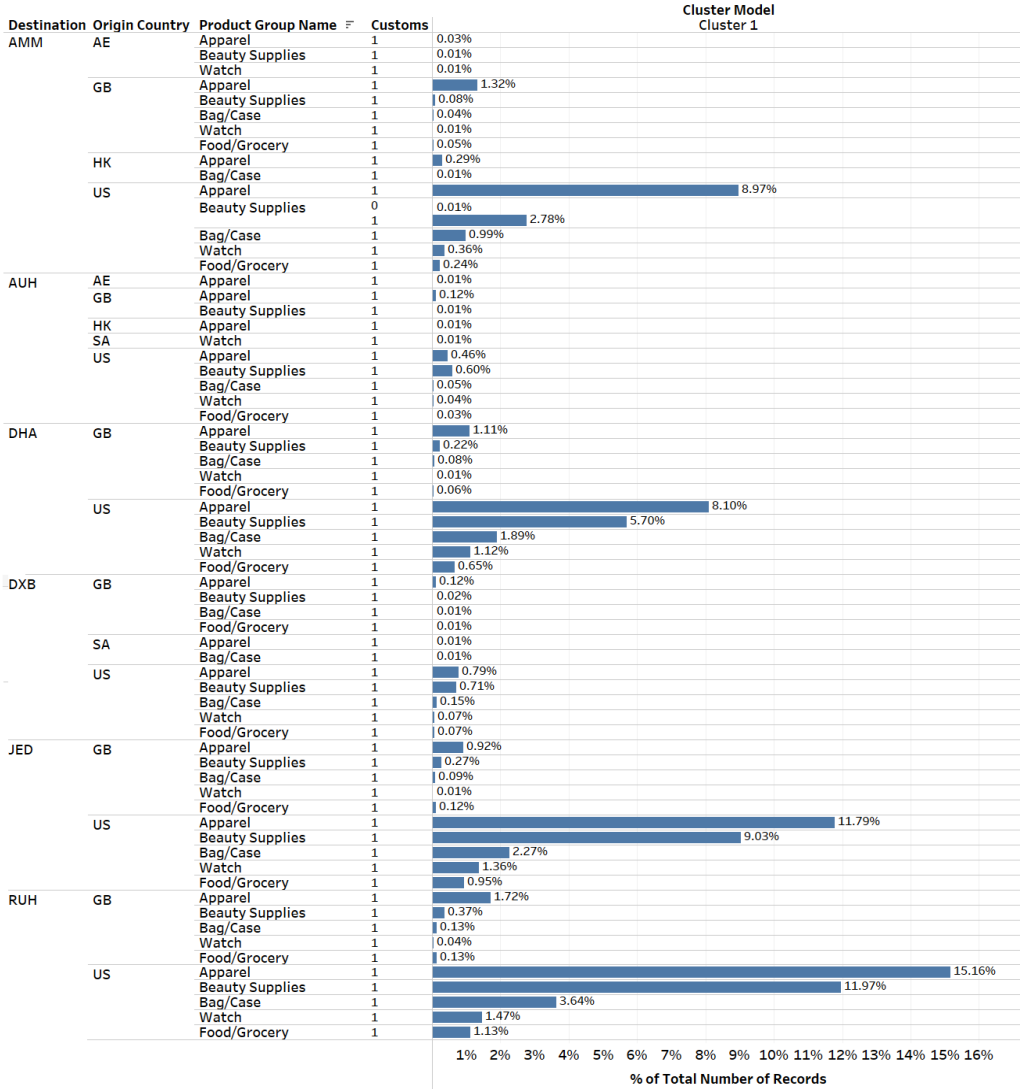

Figure 7: The distribution of the highest number of transactions in cluster-1 in detail of the 5-cluster solution expressed in percentages.

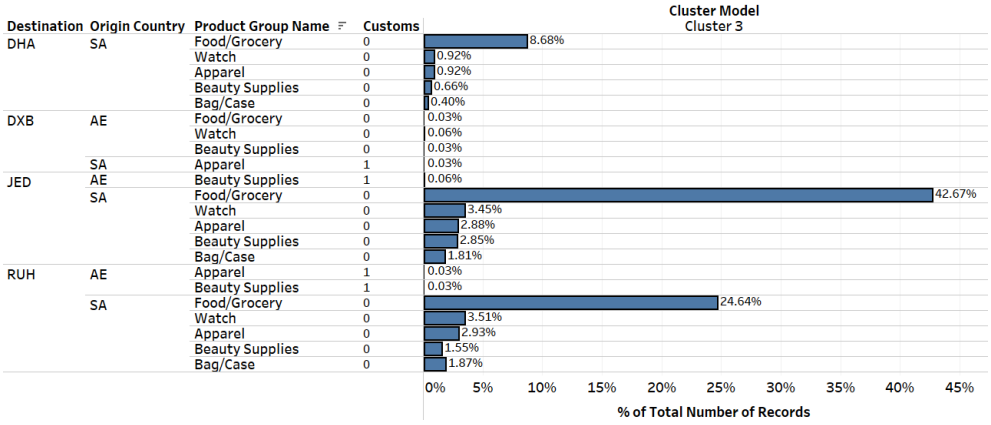

Figure 8: The distribution of the highest number of transactions in cluster-3 in detail of the 5-cluster solution expressed in percentages.

10. Results of the Semi-structured interviews

Appendix A and B respectively enumerate the questions asked during the semi-structured interviews with the key managers of both the e-company partners of our logistics company and with the manager of the logistics company.

10.1 Results for the E-companies interviews

In these companies we summarised the answers of the four companies involved in the semi-structured interviews.

Market data are of extreme importance to each of these companies in order to know their customers and the market better and to assess their market position compared to competitors in a more detailed way (more specifically with regards to market share (company 2) and building a customer data base (company 4). All four managers expect market data to help them in developing and offering customer centric solutions (company 1) and complementary services (company 2 and 4). The hope is that this will be beneficial for both their company (company 3 and 4) and customers (company 1 and 3).

Since the managers told us that they actually only possess customer characteristic data (companies 2,3 and 4) and transaction data (all companies), a lot of data are lacking. This starts with data on market shares (company 1) and customer preferences on products and sources (company 2 and 3) and ends with bases for customer clustering and segmentation like income, location and physiological trends (company 4). In general managers thus lack deeper insight in and complementary data on customers. The simple data they actually possess are clearly insufficient.

These data are mainly gathered via transactional data. Two companies possess a specific department or some people responsible for this. One company (company 4) gathers the data in a more creative way, namely by asking customers to upload their data on an online blog, posted on their website and actually attached a customer referral marketing campaign to it by adding a gift when new contacts were suggested.

However, companies 1,2 and 4 also purchase some customer contact data from third party sources, but never data on preferences and e-company sources. This has some advantages: it helps them in starting new campaigns (company 1 and 2) or in developing customer profiles (in the hope of segmenting more efficiently – company 4). Selection of these third parties is done on the basis of reliability (quality, delivery time, service – company 4), integrity (company 1 and 4), security (company 1 and 4).

Later in the enquiry the same characteristics were also mentioned for the data companies would eventually be willing to buy and for the companies they would buy them from. On this last element, also the fact that the third party should never be a competitor (company 3) was explicitly mentioned. Three of the four companies are indeed interested in getting the lacking insights in their market, customers and trends by buying the information from a third party. Only company 4 indicated that investing in market research also offers advantages over buying data from third party sources. It helps better in getting very specific information and learn from the experience by being capable of understanding the market needs better. If data would have to be bought from third party sources, their format should be capable of offering additional insights easily, that is in tabular of graphical format, dashboards being explicitly mentioned by (company 2).

This overview tells us that in general data are very valuable for the e-companies in developing a more effective and efficient marketing. Moreover, they are willing to buy the lacking data from third party sources taking into consideration the characteristics of the data mentioned in the literature are fulfilled. This is a first indication that data monetisation on logistics data could be very well possible. With regards to the type of data monetisation, selling is not the only way in which the logistics data could be monetised as our respondents clearly mentioned that deeper insights, generated through visualization and further analysis via eventual dashboards would be very useful.

The last question wanted to test whether the data generated from the actual data base we are using through visualization and further analysis would be extremely useful for our respondents. We presented some of our results as data, our respondents could potentially buy. They all four were very excited by

the possibility to buy these data. They were interested in all formats we offered them: customer preferences, payment, customs service data, market and competitor data and clustering results. It proves that carefully managed and presented data are monetisable with these respondents in an actual test environment, maybe even more than the general answers before mentioned. All respondents immediately came up with the benefits they could make from all these data, such as initiating campaigns (companies 1,2,3 and 4), predicting customer reactions to products (companies 2, 3 and 4), predicting customs application (companies 1, 2 and 3), focusing on areas competitors (companies 1,2 and 3), detailed market segmentation and clustering (companies 1,3 and 4) and offering complementary services (companies 1, 2 and 3). This is a sign of the value of the data base the logistics company actually possesses and of the necessity to visualize and analyse them deeper before monetising their data. They could from this small experiment be considered as a source of income.

Six months after we conducted our enquiry, we contacted the e-commerce companies again to check whether they had continued talking to the logistics company about the data. Three of them confirmed having bought the data, with the fourth one still in the process of negotiating about it. This proves the monetisability of the data, although we could not yet verify which value the data exactly represented as we did not get access to their financial statements, which would have clarified the impact of the use of the data on their marketing and sales efforts.

10.2 Results for the logistics company interview

Contrary to the enquiry with the e-companies, we showed our respondent in the logistics company the data they possess in the format we developed by visualizing and analysing them.

Our respondent told us that actually no data at all are used by, shared with or sold to partners or customers at the moment, although many of their partners indicated to be interested in them, particularly in the data on the number of returns based on a successful 'COD' or not, and most importantly on detailed

customs values. Therefore, our respondent really believes that making some of these data available to their partners and making them easy to be accessible, would activate business and improve the quality of their services.

It would offer the company new market opportunities and gain them "easy money". Engaging in making these data available and selling them would interest customers as most of the data, in the eyes of our respondent, would cover missing market information for their customers/partners. It would be a win-win situation for both parties as it would help the partners in solving unaddressed issues. Our respondent considers this to be a realistic extra service to be offered given the quality of the data and the absent availability on the partners' side. Future services could also include investing in solutions for the challenge of the rising number of returns.

The logistics company also thinks the data we prepared for them would also be extremely valuable to their own business and can be monetised. The reasons are more specifically:

1- The customer data contain detailed information about the purchases and the commodity of each purchase. This would be of huge help in initiating campaigns to target customers based on their interests.
2- Preferred payment info and detailed customs values per destination would, if shared within the data, give the company estimates and forecasts about how much returns can be expected, and why certain destinations are having drops in volumes.
3- Data about other e-commerce competitors will give the company insight into the current market situation. It would indicate whether they are doing well compared to competitors. This would let them analyse their performance better and signal improvement areas.
4- The detailed clustering results were considered to be the most interesting ones. They could give great flexibility in visualizing the data and manipulating needed dimensions as per our interests.

Given these answers, we can conclude that the logistics company indeed sees the opportunity of data monetisation with the data base they possess, but more specifically if these data would be transformed through deeper analysis and visualization so as to make it easier for them and for customers/ partners to gain deeper insight. They consequently realise that probably the category of

offering services which provide deeper knowledge is more interesting than just selling the data in terms of the categorisation theory mentioned before. Therefore, the second and third data monetisation techniques are preferred to monetise the data. They are: Providing insights or analyses and creating platform for new services. This is very congruent with our conclusions for the e-companies.

Our respondent developed a business model for monetising the data using the business model canvas proposed by Osterwalder et al. (Osterwalder, Pigneur et al. 2010). Table 5 shows the results.

From the business model canvas we can read that the data monetisation efforts will provide the company with the possibility to offer improved and enlarged service packages to their partners. This will help them in increasing sales and improving their market share as their knowledge of the market will be enhanced. Our logistics company mentioned that from the cost perspective, the company already has R&D and Marketing departments. They can treat this new market easily. Consequently, a monetisation project will not cost much since it does not need special or new IT infrastructure. The revenue is based on sharing the data with their partners as an extra service. This revenue could be collected directly.

The conditions relating to the process in the company mentioned before can be drawn from the business model canvas. For instance, connecting and sharing the data with the e-companies and partners accomplished using the key activities and channels segments. This is relevance to the technology factor. Whereas, the selling mode chosen by the company is double: it contains both providing insights or analyses and creating a platform for new services. This is related to both business and execution factors.

Table 5: Data monetisation business model canvas.

<table>
<tr><th colspan="5">Business Model Canvas</th></tr>
<tr><td>Key Partners</td><td>Key Activities</td><td>Value Propositions</td><td>Customer Relationships</td><td>Customer Segments</td></tr>
<tr><td rowspan="3">E-commerce companies / partners</td><td>R&D (Research and Development), building customers database on the base of products interest</td><td rowspan="3">Data monetisation, improved marketing activities, increased sales and market share</td><td>Analytics to help partners. Trust in the data integrity and security. Partnership to serve one goal</td><td rowspan="3">Partners, e-commerce companies</td></tr>
<tr><td>Key Resources</td><td>Channels</td></tr>
<tr><td>Employees, Data, IT infrastructure</td><td>Visits, meetings, marketing campaigns, conference calls</td></tr>
<tr><td colspan="3">Cost Structure</td><td colspan="2">Revenue Structure</td></tr>
<tr><td colspan="3">R&D costs, Marketing costs</td><td colspan="2">Service sales, Data share revenue from partners</td></tr>
</table>

11. Discussion and Conclusion

Visual analytics gives the logistic companies a clearer image allowing them to understand their market in more detail and adapt their marketing accordingly. All customers can be grouped in different ways, such as by 'Products', 'E-companies', 'Destination cities' and so on. The results vary according to the different variables used in combination with each other which is proven by our results of the market analyses. Moreover, logistics companies would benefit from the results of the e-companies' market analyses and better understand the competitiveness of the market of e-companies, eventually looking at ways to develop a more unique selling proposition (for instance via e-cards for special occasions or document safety services) as this seems to be a highly competitive market. It can therefore improve the market image and sales volume of the e-companies.

Every e-commerce company or customer can therefore indeed be assessed by logistics companies according to the results of the visualization. Knowing the distribution pattern of the shipments according to 'Product types', 'Customers', 'Cities' and so on is indeed highly valuable. It can direct the attention of the logistic companies to specific patterns which allows them to better target their marketing efforts. This type of information is highly significant to all participating partners.

Visual analytics, while being highly useful to manage the transactions, is however only a first step in answering our research question. For instance, it could help the logistics companies to monetise their data and selling it to other e-commerce companies.

To be able to potential monetise the data, the data should not only be well presented, but also adhere to the characteristics of monetisability. Are these two elements present?

With regards to these four criteria of data quality we can be fairly sure that they are present. The data are reliable and accurate since they are taken from the real dataset of logistics transactions by the logistics company. They are relevant to the customer companies as the e-companies are their partners and are all situated in the same sector and region. The encryption of the name of the

customers and e-companies makes the data anonymous, moreover, the results are well segmented, as the clustering results show. Therefore, all four criteria for data quality previously mentioned are fulfilled.

We can indeed say that the data are clearly ordered and presented in such a way that some of the data have become interesting to many partners.

In order to find out whether the company has the right mind-set to really market the data, the question was investigated by looking deeper into the business model the company could use in doing so by having semi-structured interviews with some of the partner's managers involved and with a manager of the logistics company. The results prove that the partner companies consider the data to be valuable enough to be invested into and the logistics company confirms this idea. Moreover, the logistic company agreed with its E-partners on what we proposed, namely that the second and third data monetisation techniques were the best to monetise our data, namely providing insights or analyses and creating a platform for new services.

All this evidence points to the fact that our research question "**What is the possibility of data to be monetised in the logistics sector?"** can be answered positively given the right visual analytics and business model used. The fact that three of the four contacted companies have since our investigation bought the data proves the value of them and that they are monetisable.

Bibliography

Bataineh, A. S., R. Mizouni, M. E. Barachi and J. Bentahar (2016). "Monetizing Personal Data: A Two-Sided Market Approach." Procedia Computer Science 83: 472-479.

Bélanger, F. and R. E. Crossler (2011). "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems." MIS Quarterly 35(4): 1017-1041.

Bonneau, V. (2015). "Data Monetisation: Opportunities beyond OTT: finance, retail, telecom and connected objects." Communications & Strategies 97: 123-126,151.

Carmona, C. J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus and S. García (2012). "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com." Expert Systems with Applications 39(12): 11243-11249.

Carnelley, G. M. H. S. P. (2018). Update of the European Data Market: Data monetisation. Luxembourg.

Catarina Arnaut, M. P., Elizabeth Scaria, Arnaud Berghmans,Sophie Leconte (2018). Study on data sharing between companies in Europe, everis.

Company, M. (2017). Fueling growth through data monetization

Derwisch, S. (2019). Data Monetization – Use Cases, Implementation and Added Value. Germany, The Business Application Research Center (BARC).

FRED, J. (2017). Data Monetization – How an Organization Can Generate Revenue with Data? , Tampere University of Technology.

Huang, D., M. Tory, B. A. Aseniero, L. Bartram, S. Bateman, S. Carpendale, A. Tang and R. Woodbury (2015). "Personal Visualization and Personal Visual Analytics." IEEE Transactions on Visualization and Computer Graphics 21(3): 420-433.

Kotu, V. and B. Deshpande (2019). Chapter 7 - Clustering. Data Science (Second Edition). V. Kotu and B. Deshpande, Morgan Kaufmann: 221-261.

Laitila, M. (2017). Data monetization: Utilizing data as an asset to generate new revenues for firms, Aalto University.

Laudon, K. C. (1996). "Markets and privacy." Commun. ACM 39(9): 92-104.

Li, X.-B. and S. Raghunathan (2014). "Pricing and disseminating customer data with privacy awareness." Decision support systems 59: 63-73.

Liu, Y. H., Y. Ren and R. Dew (2009). "Monetising user generated content using data mining techniques." AusDM.

Moore, S. (2015). "How to Monetize Your Customer Data."

Najjar, M. and W. Kettinger (2014). "Data Monetization: Lessons from a Retailer's Journey." MIS Quarterly Executive 12.

Osterwalder, A., Y. Pigneur and T. Clark (2010). Business model generation : a handbook for visionaries, game changers, and challengers. Hoboken, NJ, Wiley.

Pak Chung Wong, J. T. (2004). "Visual Analytics." IEEE Computer Graphics and Applications 24(05).

Papamichail, G. P. and D. P. Papamichail (2007). "The k-means range algorithm for personalized data clustering in e-commerce." European Journal of Operational Research 177(3): 1400-1408.

Perrons, R. K. and J. W. Jensen (2015). "Data as an asset: What the oil and gas sector can learn from other industries about "Big Data"." Energy Policy 81: 117-121.

platform, L. s. d. m. (2018). "How to Monetize Your Data." How to Monetize Your Data.

Qabbaah, H., G. Sammour and K. Vanhoof. (2019). Using K-Means Clustering and Data Visualization for Monetizing logistics Data. 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE.

Sadiku, M., A. Shadare, S. Musa, C. Akujuobi and R. Perry (2016). "DATA VISUALIZATION." International Journal of Engineering Research and Advanced Technology (IJERAT) 12: 2454-6135.

Serge Findling, M. S., Lynne Schneider, Dan Vesset (2018) "IDC PlanScape: Data Monetization."

Thomas, L. D. W. and A. Leiponen (2016). "Big data commercialization." IEEE Engineering Management Review 44(2): 74-90.

Xu, G., X. Qiu, M. Fang, X. Kou and Y. Yu (2019). "Data-driven operational risk analysis in E-Commerce Logistics." Advanced Engineering Informatics 40: 29-35.

Appendix A E-commerce companies interview

How valuable are market data for you ?

- How valuable are market data for you?

Which data about your market do you possess and use ?

Which data are lacking ?

Which of the data you use do you gather yourselves ?

- How do you do so ?

Which of the data you use are you currently buying from a third source?

- Which advantages does this have for you ?
- Which criteria do you use to select such a third source ?

Which ways would you like to use to obtain the lacking data ?

- Would you prefer to invest in further research yourselves or be willing to buy these data from a third source ?
- If you would be buying these data from a third source, what characteristics should the data have?
- How should they be presented ?
- Which characteristics should this third source have ?

We are now going to present some types of data that logistics partners could eventually present and offer to you?

please indicate whether you are interested to buy these data, if yes (what makes these data valuable for your business?

Appendix B Our logistics company interview

We are presenting you some of the data you possess.

Do some of your partners/customers actually use some of your data ?

- Which data are involved ?
- How many partners/customers are interested ? (many, some, only a few)

Do you actually share data with partners/customers?

- Which data are involved ?
- How many partners/customers are interested ? (many, some, only a few)

Did you already get questions by partners about the data you possess and that could be interesting to them ?

- If so, About which data did these questions handle?
- How many partners/customers asked about this ? (many, some, a few,..)

What is in your eyes the value of the data shared/used with your partners for them?

Why are they willing to engage in these activities according to you ?

- What are the advantages for them?

Do you currently sell data to partners/customers or provide them for free?

- If yes, which ones? At which value?
- In which way do these data contribute to your revenue?
- Do you think this could be improved ? In which way ?
- Which processes inside your company should/could be improved to be more effective/successful in this respect ?
- If no, how realistic do you think it would be to provide your partners/customers with these data for a certain price (selling them) ?
- How realistic would providing extra services on the basis of these data can be?

- Would offering new services be a good form of getting revenue from them for your partners/customers? Why/why not?
- Which services could be involved ?

Do you think these data can either be sold/or-and/offered as an additional service to your partners? If yes. What makes these data valuable to your business? What is your business model to monetise these data?

Authors' Information

Hamzah Qabbaah *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B6b, Campus Diepenbeek, B-3590 Diepenbeek, Limburg,Belgium e-mail: hamzah.qabbaah@uhasselt.be*

Major Fields of Scientific Research: Data mining, Digital marketing, Business Intelligence, Machine Learning, Knowledge Management

George Sammour *- Director of Quality Assurance and Accreditation Centre, Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan. e-mail: George.Sammour@psut.edu.jo*

Major Fields of Scientific Research: Data mining, Digital learning, lifelong learning, professional development

Koen Vanhoof *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek, Limburg,Belgium e-mail:koen.vanhoof@uhasselt.be*
Major Fields of Scientific Research: data mining, knowledge retrieval

ON LOGICAL-COMBINATORIAL SUPERVISED REINFORCEMENT LEARNING[1]

Levon Aslanyan, Vladimir Ryazanov, Hasmik Sahakyan

***Abstract**: In this paper we consider a novel and important postulation in area of pattern recognition, where instead of the accurate object classification into the classes by the learning set, the objective is to assign all objects to the same, the so-called, "normal" class. We are given a learning set L; among the classes there is one called "normal" class K_0, and l "deviated" classes $K_1, K_2, \dots, K_l$ from some environment K. The learning process is dynamic in recurrent "classification, action" format in the following way: a certain action/function $\mathcal{A}_i$ is attached to each of the "deviated" classes K_i, such that applied to an arbitrary object $x \in K_i$, the action delivers its update $\mathcal{A}_i(x)$, keeping it in the same environment K. As a result, $\mathcal{A}_i(x)$ may be classified either to one of the deviated classes (included the same class K_i), or to the "normal" class K_0. The goal is in constructing a classification algorithm $\mathfrak{A}$ that applied repeatedly (small number of times) to the objects of L, moves the objects (correspondingly, the elements of K) to the "normal" class. In this way, the static recognition is transferred to a dynamic domain.*

This paper is a discussion on the problem, its theoretical postulations, possible use cases, and advantages of using logical-combinatorial approaches in solving dynamic recognition problems. Some light relation to the topics like reinforcement learning and recurrent neural networks will be taken into account

[1] The ideas of this article are presented in the proposal "Machine Learning Theory and Converging Use Cases for Linguistics and Genomics" submitted to SC of MESCS Armenia

***Keywords**: classification, logical-combinatorial approach, supervised reinforcement learning*

***ITHEA Keywords**: I.2 ARTIFICIAL INTELLIGENCE.*

Introduction and Problem Statement

The typical case pattern recognition problem considers n features, disjoint classes $K_1, K_2, \dots, K_l$ from some environment K, and an m object learning set $L = \{x_1, x_2, \dots, x_m\}$, where $L \cap K_i, i = 1,2, \dots, l$ is the share of the i-th class in the learning set. The goal is to create a classification algorithm $\mathfrak{A}$ based on the learning set, which classifies objects in the environment K as accurate, as possible. Additional information about the classes and classification is a benefit.

We consider a principally different version of the pattern recognition problem, where it is assumed that one of the given classes is "normal", let it be denoted as K_0, and all the other classes are "deviated classes". Also, we are given a finite set $\mathcal{A}$ of actions/functions a, that being applied to the objects $x \in K$ deliver their functional updates $a(x)$, keeping them in the same environment K. In the simplest case we assume that a certain action $\mathrm{a}_i \in \mathcal{A}$ is attached to every "deviated" class K_i (the i-th class action); and being applied to an arbitrary object $x \in K_i$, delivers its update $\mathrm{a}_i(x)$. $\mathrm{a}_i(x)$ may be allocated to anyone of the classes and it is not necessary that this is a unique class for all objects of K_i. The goal is in constructing a classification algorithm $\mathfrak{A}$, that applied repeatedly (small number of times) to the objects of L (correspondingly, the elements of K) moves these objects to the "normal" class.

Thus, the process is as follows: Algorithm $\mathfrak{A}$ is applied repeatedly to the elements of learning set L and their updates by the set of class actions. If after a current k-th repetition/application of the algorithm there still remains an object $x \in L$, or an object appeared during the process, which is classified not to the class K_0 (instead, it is classified, say, to some deviated class K_i), then the action

a_i, attached to the class K_i, is applied on x at the next $(k+1)$-th repetition of $\mathfrak{A}$, updating the learning set labels in this way.

Consider one application scenario of medical domain - Dynamic correction of the patient's treatment regime [Zhang, 2019]. Here, "classification operation" means that the current diagnose is obtained by the medical doctor for any object of classes $1 \div k$. There is no reason to apply classification to the "normal" class because its elements represent the healthy cases. Recall, that each of the classes K_i is $1-1$ related to their actions a_i, and in this case, a_i is the treatment action for class K_i. It is evident, that the overall goal is to bring the patients, after several treatment stages, to the "normal" class. Two different subcases of this use case will be considered. At first we suppose that the records and observations of only one particular doctor are available. In this case we aim at estimating the effectiveness of the diagnostic approaches of the doctor. In second scenario we suppose that we are given a larger information of a set of doctors and we try to determine the optimal way of diagnoses to achieve the best allocation result to the "normal" class.

In algorithmic point of view this is an inverse-recognition-problem. Ordinary recognition aims at mimics of the one-step classification actions. Here, for an algorithm that we apply recurrently, we need to guess all ancestors that will be mapped onto the predefined class. Moreover, it is necessary to generate an algorithm with the set of ancestors larger than the learning set.

Scenarios and Problem Definition

As we mentioned, two main scenarios and corresponding problems will be considered/highlighted here:

(1) **Scenario 1**: basic available information of this scenario is given in the form of a learning set L of a classification problem. Although the class actions are automatically applied to the elements of the deviated classes, and each reapplication of the algorithm may work with the updated objects, however, we are given neither this information, nor the updates themselves. We suppose

only, that empirically it is accepted/supposed that the set L is obtained/recorded in a practice by a witness, in form of object-class-label, and the objects of K tend to be classified to the class K_0 in a few repeated applications of the algorithm $\mathfrak{A}$, but this needs to be verified.

In its complete for the set L is a data flow. Considered objects x have their identifiers I_x which is many-to-one mapping. x, after operated by the algorithm $\mathfrak{A}$, changes its time stamp. Initial time stamp is the time t_0 of the first appearance in the algorithm $\mathfrak{A}$. After classification and action applied, x accepts the modified value $x^{(1)}$ and the new time stamp t_1 with $t_0 < t_1$. In this way objects travel through the classes forming the so called traces, $t_0, t_1, \dots, t_k$. The basic objective is to insure, that the end points of traces belong to the class "normal". In this Scenario we have a bystander, witness, who cannot see the timestamp and identifiers. In these limited information the problem formed will try to verify whether the strategy of algorithm $\mathfrak{A}$ is supportive to classification to the class "normal".

Problem 1: assess the compliance and validation of the empirical classification algorithm $\mathfrak{A}$ into the class "normal" based on the learning set L.

(2) **Scenario 2**: the learning set L is updated after each reapplication of the classification algorithm, according to the class actions results/updates.

Problem 2: synthesis of an optimized classification algorithm $\mathfrak{A}$ according to the extension of the learning set L, which includes also additional information on the updated objects.

Problem 3: choice of optimal/relevant class actions in both scenarios.

Proposed Methods and Solutions

Logical-combinatorial model of the pattern recognition

Here we bring basic definitions from the logic-combinatorial pattern recognition theory. This theory will be used in solving the Problems 1-3. Consider a typical case recognition problem with n features, l disjoint classes $K_1, K_2, \dots, K_l$ from an environment K and an m object learning set $L = \{x_1, x_2, \dots, x_m\}$. $L_i = L \cap K_i, i = 1,2, \dots, l$ denotes the share of the i-th class of the learning set, that we suppose, is not empty. Objects are identical to their descriptions in the form of a vector of feature values: $x = (x_1, x_2, \dots, x_n)$. For simplicity, we assume that $x_i \in R$, $i = 1,2, \cdots, n$.

Let us define the following set of elementary predicates, parametrically dependent on support sets $\omega_1, \omega_2 \subseteq \{1,2, \dots, n\}, |\omega_1| = k_1, |\omega_2| = k_2$ and vectors $c_1 \in R^{k_1}$ and $c_2 \in R^{k_2}$. Below we use the notation $(x \le a) = \begin{cases} 1, x \le a, \\ 0, otherwise \end{cases}$.

Definition 1 [Ryazanov, 2007]

The predicate $P^{\omega_1,k_1,\omega_2,k_2}(\boldsymbol{x}) = \bigwedge_{j\in\omega_1}(c_{1,j} \le x_j) \bigwedge_{j\in\omega_2}(x_j \le c_{2,j})$ is called a logical dependency (LD, geometrically a parallelotope) of the class K_i, if

1. $\exists x_t \in L_i$: $P^{\omega_1,k_1,\omega_2,k_2}(x_t) = 1$,

2. $\forall x_t \notin L_i$: $P^{\omega_1,k_1,\omega_2,k_2}(x_t) = 0$,

3. $P^{\omega_1,k_1,\omega_2,k_2}(x) = extr(F(P^{\omega_1,k_1,\omega_2,k_2}(x)))$,

where F is the predicate quality criterion.

It is clear that the defined predicate geometrically presents a parallelotope; and then the function F requires to find local maximization of LDs in the domain. We will denote the set of all LDs of the i-th class of the given problem by $P_{?_i}$, and the set of all LDs of all classes by P_L.

The predicate, satisfying only the first two constraints, is called admissible. We also consider the approximate predicates with limited violates of the condition 2.

LD is the base element of the logic-combinatorial pattern recognition (LCPR) theory. The initial idea with LD appeared in [Dmitriev, 1966]. The multi-parametric voting algorithms over the LD were introduced in [Zhuravlev, 1971]. [Aslanyan, 1975] obtained a complete analytics for LD with the binary features. Here the predicates are maximal intervals/subcubes of the partially defined Boolean function, and the set of predicates is given by the reduced disjunctive normal forms.

Definition 2 [Zhuravlev, 1998]

LCPR similarity measure of an object of recognition x, and a class K_i is:

$$\Gamma_i(x) = \frac{1}{|P_L|}\sum_{P^{\omega_1,k_1,\omega_2,k_2}\in P_L} P^{\omega_1,k_1,\omega_2,k_2}(x).$$

In short description, the LCPR stands out by:

- effective measure of similarity,
- proven separation of classes,
- multi-parametric optimization over large sets of recognition algorithms,
- correction of sets of algorithms providing correct recognition for all objects recognized by at least one individual recognizers, and other properties.

Advantages of using the LCPR in solving the dynamic recognition problems

Consider the Scenario 1.

In a general recognition algorithm $\mathfrak{A}$ by the learning set L there is no visible idea how to follow with repeated classifications. However, the situation is different

with the LCPR, because here it is possible to apply a backward reconstruction procedure of logical dependencies. At first, the set of LD for the class K_0 is constructed by L. As it was mentioned, this is a set of parallelotopes in R^n. We suppose that all elements covered by these LD create a new artificial class K_*, and one may now construct LDs defined by this class and by L. The Cartesian multiplication of the previous stage LDs, - is the way of creating new LDs. Continuing the growing process of LDs, in parallel, we compare the covered volume of the object space with the size of $\boldsymbol{L}$.

Implementation of this technique is not straightforward, it needs the knowledge gained on LCPR, as well as development of new approximate parallelotope-set type coverage approaches, to keep the appearing complexities tractable.

It is worth mentioning that LCPR with LD provides the partial geometrical data structure, that helps not only with complexity controls, but also provides interpretability of results; and this is the known comparative benefit of all LCPR approaches.

Linkage graph model

In the Scenario 2 the learning set L is updated/extended after each reapplication of the classification algorithm. In this case the object ID is recorded in all steps that provides a follow up mechanism through the recurrent classification process. Let $x \in L$, $x \in K_i$, and let some empirical treatment of x be known. That is, x is classified to the class K_j; after that, action a_j (the j-th class action) is applied, and as a result x is modified into y: $\mathrm{a}_j(x) = y$. In this way, chains are appearing in the course of repeated classifications, and some of this chains lead to the class K_0.

In a formal description, the learning set L is represented by a linkage graph G, with the vertex set V corresponding to the learning set elements, and with directed edges E, labeled by actions, connecting pairs of learning elements. An edge may have a weight or may not have. In this manner, the graph G provides a valuable information for checking the model validity, and obtaining a realistic information about the applied problems. The graph-theoretical problems that

appear here helping to check the system, are well investigated theoretically; while its analytics through the sparse symmetric diagonally dominant matrix computational theory, - will give acceptable implementation in algorithms and software.

Multi-criteria optimization problem

The problem of choosing optimal class actions, leads to a multi-criteria optimization, and in this regard, the multi-layered logical dependencies need to be investigated.

Inverse Recognition Problem

The mentioned problems, in their general form, refer to the conceptual direction of the machine learning known as Reinforcement Learning (RL). The goal of RL is to create an optimal acting agent for successful interacting with the environment. The problem formulated above is a very specific case of RL. Action is learned to classify all objects to the unique "normal" class. So, when class label is different from "normal", the action gets penalty. This approach can also be presented in a form of a recurrent neural network model. A weaker relation is with the known *inverse classification model* which is analysis the features space, and the features groups, providing a better one-class classification. No other systematically studied and related areas are known. The mentioned technique is tightly related to the backpropagation approach. Backpropagation has a very broad scope, and the "normal" class classification discipline appears as the inverse recognition problem. One step back gives the area that will/may be mapped to the class "normal". It is to differ objects that necessarily will be classified to "normal" ($\forall$), objects that never mapped to "normal" ($\emptyset$), and others, that are classified to classes in accord to some probabilistic distributions, and the class "normal" is among these classes ($\exists$). Next step back accepts a similar picture of classification. Our goal is to

determine all objects always classified to "normal", and those will be allocated to "normal" at least one time. And of course we are interested to know the frequencies of these allocations. Our technique to achieve this information is the LCPR model and algorithms.

The LCPR domain has been introduced and investigated by our team for decades, resulting in hundreds of publications and scientific theses. Most investigated is the binary case. Here the reduced disjunctive normal form (RDNF) is the analytical basis that helps to describe these classes of objects. In the simpler case of two classes two RDNF are considered. First is for the positive Boolean function that is true on the elements of "normal" learning elements and the second is for negation of this function. [Zhuravlev, 1998] shows that intersection of these two RDNF by LCPR will correspond to (∃), while the positive intervals/subcubes will denote the (∀) pats of the learning set. The (∀) of positive ("normal" learning set) Boolean function represents the backpropagation of the "normal" class.

Nest step to back is similar to the first step. Here new intervals/subcubes will be formed, the core essence is in fact that all elements of the first step intervals/subcubes will be enlarged similarly which draws to the Cartesian degrees of the intervals/subcubes. This is probably not simple but visible and interpretable analytics to inverse recognition procedures.

Multiclass extension is not difficult. It is to consider one-to-many classifications for all classes. This brings a scheme of $l + 1$ RDNFs. The reminder is similar to the two class example. Of course this is an initial interpretation of the inverse recognition model by the use of LCPR. The studies will be continued and implemented in practice. We evaluate the first step done as important as the applied model is ultimately practical being not yet formed and studied.

Conclusion

More than 70 years of development of pattern recognition, which is now referred to by the term machine learning, has made it possible to formulate a solid set of models and technologies of both statistical and logical combinatorial nature. The variety is huge, both in the form of models and scenarios, as well as technologies and algorithms. The emergence of new tasks that have not yet been formed and not studied is not excluded. One of such tasks is the task of assigning all objects to one fixed class by several consecutive steps of recognition. An applied problem of this type may be optimization of the course of treatment in medicine. This paper considers algorithms of pattern recognition logical regularities in the context of solving the problem of assignment to the one, fixed class. Only the initial research on this problem is characterized, its connection with the concept of reinforcement learning and recurrent neural networks is indicated. Subsequent investigations of the problem will prove useful in a number of applied problems.

Bibliography

[Ryazanov, 2007] Ryazanov V. V., Logical regularities in pattern recognition (parametric approach)", Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, vol. 47, no. 10, pp. 1793-1808, 2007.

[Dmitriev, 1966] Dmitriev A.N., Zhuravlev Yu. I. and Krendelev F. P., On mathematical principles of object and phenomena classification, Discrete Analysis, Novosibirsk (In Russian), vol. 7, pp. 3-15, 1966.

[Zhuravlev, 1971] Zhuravlev Yu. I., Nikiforov V. V., Recognition algorithms based on calculation of estimates, Cybernetics, vol. 3, pp.1-11, 1971.

[Aslanyan, 1975] Aslanyan L. H., On a pattern recognition method based on the separation by the disjunctive normal forms, Kibernetika, vol. 5, pp. 103 -110, 1975.

[Zhuravlev, 1998] Zhuravlev Yu. I., Selected Scientific Works, (in Russian), Magistr, Moscow, 1998.

[Zhang, 2019] Zhang Z., Reinforcement learning in clinical medicine: a method to optimize dynamic treatment regime over time, Annals of Translational Medicine, 2019, 7(14):345.

[Gimenez-Martinez, 2001] V., Aslanyan L., Castellanos J., Riazanov V., Distribution Function as Attractors for Recurrent Neural Networks, Pattern recognition and image analysis, vol. 11.3, 2001, pp. 492-497.

[Neu, 2007] Neu G., Szepesvári C., Apprenticeship learning using inverse reinforcement learning and gradient methods, In: Proc. 23rd Conf. Uncertainty in Artificial Intelligence, (2007) pp. 295-302.

[Aggarwal, 2010] C. C. Aggarwal, C. Chen, and J. Han, The inverse classification problem, Journal of Computer Science and Technology, vol. 25, no. May, pp. 458-468, 2010.

[Sutton, 1998] Sutton R, Barto A., Re-Inforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1988

Authors' Information

Levon Aslanyan - *Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia; Head of Department of Discrete Mathematics. 1 P.Sevak str. Yerevan-0014, Armenia; e-mail: lasl@sci.am.*

Major Fields of Scientific Research: mathematical logic, discrete mathematics, mathematical theory of recognition, and artificial intelligence.

Vladimir Ryazanov – *Dorodnitsyn Computing Centre, Federal Research Center Computer Science and Control,*

Russian Academy of Sciences; Head of the Department of Methods of Classification and Data Analysis. Vavilova 44, Moscow, 119333 Russia; e-mail: rvvccas@mail.ru.

Major Fields of Scientific Research: optimization methods of recognition models, algorithms for searching and processing logical regularities by precedents, mathematical recognition models based on voting by sets of logical regularities of classes, committee synthesis of collective clustering and construction of stable solutions in clustering problems.

Hasmik Sahakyan – *Institute for Informatics and Automation Problems of the National of Science of Armenia; Scientific Secretary. 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: hsahakyan@sci.am*

Major Fields of Scientific Research: Combinatorics, Discrete tomography, Data Mining.

OPTIMIZATION PROBLEM: SYSTEMIC APPROACH

Albert Voronin, Yuriy Ziatdinov

Abstract. *A systemic approach to the problem of multi-criteria optimization allows us to combine the models of individual compromise schemes into a single holistic structure, adapting to the situation of multi-criteria decision making. The advantage of the concept of a non-linear compromise scheme is the possibility of making a multi-criteria decision formally, without direct human participation.*

Keywords: *system, optimization, multicriteria problem, utility function, scalar convolution, non-linear scheme of compromises.*

ACM Classification Keywords: *H.1 Models and Principles – H.1.1 – Systems and Information Theory; H.4.2 – Types of Systems.*

Introduction

The essence of many practical problems in different subject areas is the choice of conditions that allow the object of research in a given situation to show its best properties (optimization problem). The conditions on which the properties of the object depend are expressed quantitatively by some variables x_1, x_2, … , x_n, given in the domain of definition X and called optimization arguments. External actions r do not depend on us, but it is known that they can take their values from a compact set R. Usually it is assumed that the calculations are carried out for a given and known external action vector $r^0 \in R$, which ultimately determines the decision-making situation.

In turn, each of the properties of the object in the domain M is quantitatively described by the variable y_k, $k\epsilon[1,s]$, the value of which characterizes the quality of the object O in relation to this property (Fig. 1):

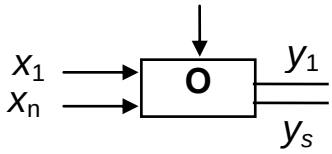

Fig. 1

In the general case, the parameters y_1, y_2, … , y_s, called the quality criteria, form the vector $y = \{y_k\}_{k=1}^{s} \epsilon M$. Its components quantify the properties of the object for a given set of optimization arguments $x = \{x_i\}_{i=1}^{n} \in X$.

Systemic Approach

The term "systemic approach" means that a real object represented as a system is described as a set of interacting components that implements a specific goal. At the same time, a finite, but ordered set of elements and relations between them is "cut out" from the variety of components of a real object. We can say that the system is a model of a real object only in the aspect of the goal that it implements. The goal, requiring for its achievement certain functions, determines through them the composition and structure of the system.

The goal isolates, outlines the contours of the system in the object. In this system (object model) only what is necessary and sufficient to achieve the goal will come from the real object. If the same object can realize several goals, then with respect to each it acts as an independent system. The systemic approach

assumes that not only the object, but also the research process itself acts as a complex system, the task of which, in particular, consists in combining in a single whole various models of the object.

Thus, with the systemic approach, the researcher receives only that information about a real object, which is necessary and sufficient to solve the task.

Optimization

If the object realizes only one goal, then the effectiveness of achieving the goal is quantitatively expressed by the single criterion of optimality y. The solution of the optimization problem involves reaching the extreme value of the criterion by choosing the set of optimization arguments.

The extremalization of the optimality criterion is often identified with the concept of goal realization, while in reality these are different concepts. We can say that the criterion and goal are correlated with each other as a model and an original with all the consequences that follow from this. This is understandable, if only because the original is usually put in line not one, but several models reflecting this or that aspect of the original. Some goals are difficult, and sometimes impossible to describe with the help of quantitative criteria. In any case, the criterion is just a surrogate of the goal. Criteria characterize the goal only indirectly, sometimes better, sometimes worse, but always approximately [1,2].

The decision of optimization problems assumes presence of some estimation of quality of work of a system from which it is possible to tell that one system works better, and another – is worse and how much. The fundamental problem of the quantitative assessment of objects and processes is that the notions of

"better" and "worse" be put in line with the concepts "more" and "less." For certainty, it is believed that, for example, "better" means "less".

According to S. Stevens, if the description opens the way for measurement, then the discussions are completely replaced by calculations. In application to our problems, this means that if there are reasonable quantitative criteria for the quality of a complex system, then its study can be carried out through a formalized mathematical apparatus. Otherwise, subjective assessments, multivalued interpretations and arbitrary decisions are inevitable.

The function $y = f(x)$ relates the quality criterion to the optimization arguments. In estimation problems, the function $f(x)$ is called the *evaluation* function, and in optimization problems it is called the *objective* function. With some reservations, the optimization problem is formulated as finding such a combination of arguments from their domain, in which the objective function acquires an extreme value:

$$x^* = \arg \underset{\substack{x \in X \\ y \in M}}{\operatorname{extr}} f(x) \Big|_{r^o \in R}$$

If "better" means "less", then in practice, for fixed $r^o \in R$ and guaranteed $y \in M$, expression

$$x^* = \arg \min_{x \in X} f(x)$$

is applied.

Multicriteria Problems

A complex object of research can not be characterized by any one (for example, the most "important" or "typical") attribute. When describing it, many inseparable properties must be taken into account simultaneously. In other words, to study complex objects, a modern systemic approach requires the involvement of the entire spectrum of their properties. A complex object and any fragment of it must be viewed not in isolation, but in numerous contradictory interactions and, importantly, in various possible situations.

Complex systems, being in different conditions (situations, modes), reveal different system properties, including those that are incompatible with none of the other situations separately. In their study, an approach is used that consists in the creation and simultaneous coexistence of not one but a set of theoretical models of the same phenomenon, some of which conceptually contradict each other. However, not one can be neglected, since each characterizes some property of the phenomenon under study and neither can be accepted as a single one, since it does not express the complete complex of its properties. It is interesting to compare what was said with the principle of complementarity introduced into science by Niels Bohr: "... To reproduce the integrity of the phenomenon it is necessary to apply mutually exclusive "additional" classes of concepts, each of which can be used in its own special condition, but only taken together, exhaust all the determinable information".

Multiple properties of a complex system in a given situation of its functioning are quantitatively estimated by corresponding partial criteria. In different situations, the rank of "the most important" acquire different properties and, accordingly, different partial criteria. Thus, mutually exclusive "additional" classes of concepts, in which the role individual theoretical models are presented, are characterized by conflicting partial criteria, each of which is applicable in its own special condition. And only a complete set of partial criteria (vector criterion)

makes it possible to adequately assess the functioning of a complex system as a manifestation of the contradictory unity of all its properties. Therefore, it can be assumed that multicriteriality is the embodiment of the principle of complementarity in the methodology of research of the complex systems.

However, this possibility is only a necessary, but not sufficient, condition for vector estimation of the entire system as a whole. Indeed, let it be known the numerical values of all the partial criteria of the system. Does this mean that we, knowing these values, can evaluate the effectiveness of the system as a whole? No we can not.

It is appropriate to recall the old Indian parable about the blinds that got to know the elephant. One touched the trunk and decided that the elephant was like a snake. The second picked up an ear and said that the elephant reminds him of a sheet. The third felt a foot and said that the elephant is a pillar.

For a holistic assessment, it is necessary to rise to the next level, i.e. carry out the act of composing the criteria. Let us compare this with Kurt Gödel's incompleteness theorem: "... In any sufficiently complex first-order theory there is an assertion that can not be proved or disproved by the means of the theory itself. But the consistency of one particular theory can be established by means of another, more powerful, second-order formal theory. But then the question arises of the consistency of this second theory, and so on." Gödel's theorem seems to be the methodological basis for composing criteria, which is a sufficient condition for vector estimation of the system as a whole.

A scalar convolution of the criteria can serve as an instrument of the composition act. Scalar convolution is a mathematical method of compressing information and quantifying its integral properties by one number.

In general, the simultaneous description of the phenomenon (object) from several sides always gives a qualitatively new, more perfect idea of the described phenomenon (object) in comparison with any "one-sided" description. So, even two flat images that form a stereo pair make up a three-dimensional image of the object, not to mention the possibilities of holography. A multi-criteria approach that gives a "stereoscopic" look at the evaluation of the functioning of the system opens up new ways for improving complex management systems and decision making. So, for a holistic perception of a complex system in different conditions of its work, it is necessary to apply a multi-criteria approach.

In practical problems, a real object usually implements not one, but several goals and, accordingly, is characterized by several partial criteria of efficiency (quality). Let's pay attention to the fact that quality criteria, as a rule, are contradictory. The art of the researcher consists in the systemic linking models characterized by contradictory indicators. Thus, Jean Colbert (Minister of Louis XIV) in 1665 said: "The art of taxation is to get a maximum number of feathers while digging a goose, with a minimal hiss." At the systemic approach there is a task which consists in connection in a single whole of various models of object. The problem is solved by applying the act of criteria composition.

For the systemic linking in multicriteria problems, the scalar convolution of the particular criteria $Y=f[y(x)]$, where y is no longer a scalar, but an s-dimensional vector of the criteria $y = \{y_k\}_{k=1}^{s}$, is used as the objective (or evaluation) function instead of $y=f(x)$. Scalar convolution acts as an instrument of the act of composing criteria.

In the notion of optimality, in addition to the criteria, limitations $x \in X$ on the optimization arguments as well as $y \in M$ on the efficiency of the solution play an equally important role. Even small changes can significantly affect the solution. And very serious consequences can be obtained by removing certain

restrictions and adding others with the same system of criteria. There is a great danger in the optimization of complex systems, as N. Wiener pointed out in his first publications on cybernetics. The fact is that, without setting all the necessary restrictions, we can, simultaneously with the extremization of the objective function, obtain unforeseen and undesirable accompanying effects.

To illustrate this N. Wiener liked to bring an English fairy tale about a monkey's foot. The owner of this talisman could fulfill any desire with its help. When he once wished to receive a large sum of money, it turned out that for this he paid the life of his beloved son. We will agree that it is often very difficult, and sometimes it is simply impossible to foresee in advance all the consequences of adopting multi-criteria decisions.

The idea of N. Wiener that in complex systems, we are fundamentally unable to determine in advance all the conditions and limitations that guarantee the absence of undesirable optimization effects, allowed him to make a gloomy assumption about the catastrophic consequences of cybernation of society.

Nevertheless, from the standpoint of system analysis, the attitude to optimization can be formulated as follows: it is a powerful means of increasing efficiency, but it should be used more cautiously as the complexity of the problem increases.

We formulate the formulation of the multicriteria optimization problem in a fairly general form.

Formulation of the Problem

A set of possible solutions $X \subset E^n$ consisting of vectors $x=\{x_i\}_{i=1}^{n}$ of n-dimensional Euclidean space is given. By the physical nature of the problem the vector holonomic (in static) or nonholonomic (in dynamics) connection $B(x) \leq 0$ is given. The decision is made at external influences, described by the vector r, given on the set of possible factors R.

The quality of the solution is estimated from the set of contradictory partial criteria that form the s-dimensional vector $y(x)=\{y_k(x)\}_{k=1}^{s} \subset F$, which is defined on the set X. The expression $y \subset F$ denotes the vector y belonging to the class F of admissible efficiency vectors. The partial criteria vector is bounded by the admissible domain: $y \in M$.

The situation that results from the adoption of a multi-criteria solution x under given external conditions r, is characterized by the Cartesian product $S=X \times R$.

The problem is to determine a solution $x^* \in X$, which, under given conditions, connections, and constraints, optimizes the efficiency vector $y(x)$.

This formulation is so general that, according to a famous comic expression, it can not be applied in any particular case. For the constructive solution of the task in various particular statements, it is necessary to carry out the structuring of certain concepts. To do this, we need to make additional special assumptions that help solve the following problems of vector optimization:

- determination of the range of Pareto optimal solutions;

- choice of the scheme of compromises;
- normalization of partial criteria;
- consideration of priority.

Difficulties in solving vector optimization problems are not computational, but conceptual in nature (this is not *how* to find the optimal solution, but *what* should be understood by it). Therefore, the development of a formal apparatus for solving multicriteria problems is one of the most difficult problems in the modern theory of decision-making and management. Its solution is important both in theoretical and applied terms.

Selection of the Scheme of Compromises

From the problems of vector optimization, we will pay special attention to the problem of choosing a scheme of compromises. One of the most important theses of the theory of decision-making under many criteria is that there is no best solution in some absolute sense. The decision made can be considered the best only for the person making the decision (decision maker, DM) in accordance with the goal set by him and taking into account the specific situation. The normative models for solving multicriteria problems are based on the hypothesis of the existence in the consciousness of the DM some utility function [3], measured both in nominal and in ordinal scales. The reflection of this utility function is the scheme of compromises and its model in a given situation – the scalar convolution of partial criteria $Y[y(x)]$, which allows constructively solving the problem of multicriteria optimization.

The determination of a multi-criteria solution is by its nature compromise and fundamentally based on the use of subjective information. Having received this information from the decision maker and choosing a scheme of compromises, one can move from a general vector expression to a scalar convolution of

partial criteria, which is the basis for formation a constructive apparatus for solving multicriteria problems. If the scalar convolution method is used, the mathematical model of solving the vector optimization problem is represented in the form of the extremization of the function $Y[y(x)]$. This is a scalar function that has the meaning of a scalar convolution of the vector of partial criteria, the form of which depends on the chosen scheme of compromises.

The most commonly an **additive** (linear) scalar convolution is used

$$Y[y(x)] = \sum_{k=1}^{s} a_k y_k(x),$$

where a_k are the weight coefficients determined by the decision maker, starting from his utility function in the given situation. The Laplace principle in the theory of decision-making consists in the extremization of a linear scalar convolution. The drawback (specificity) of the application of linear scalar convolution is the possibility of "compensating" one criterion at the expense of others.

Multiplicative convolution

$$Y[y(x)] = \prod_{k=1}^{s} y_k(x)$$

is free of this shortcoming. The Pascal principle is the extremization of the multiplicative scalar convolution.

Historically, Blaise Pascal's principle was first described in the work of Pensees, published in 1670. It is believed that this work laid the foundation for the whole theory of decision-making. Here are introduced two key concepts of the theory: 1) partial criteria, each of which evaluates any one side of the effectiveness of the solution and 2) the principle of optimality, i.e. rule, allowing by the values of the criteria to calculate a single numerical measure of the effectiveness of the solution (act of criteria composition).

The Pascal principle is adequate in tasks with a cumulative effect, when the effect of certain efficiency factors is, as it were, increasing or decreasing the influence of other factors. When maximizing partial criteria, the zero value of any of them completely suppresses the contribution of all others to the overall effectiveness of the solution. In the aerospace industry, this approach can be partly justified when each criterion (for example, reliability and safety) is critical and no improvement in other criteria can compensate for its low value. If at least one of the partial criteria is zero, then the global criterion is also zero.

Shortcoming of application of multiplicative scalar convolution: a very expensive and very effective system can have the same estimation as a cheap and low effective. We will compare such "weapon systems" as an atomic bomb and a slingshot, which at a low cost has some damaging effect. Guided by the multiplicative convolution, it is possible to select a slingshot for the armament of the army.

Similar to the Laplace principle, one can generalize the Pascal principle by introducing weighting coefficients:

$$Y\left[y(x)\right] = \prod_{i=1}^{s} [y_i(x)]^{a_i}.$$

Convolution according to the **Charnes-Cooper** concept. The concept of Charnes-Cooper is based on the principle of "closer to the ideal (utopian) point." In the space of criteria under given conditions and constraints, an unknown a priori vector y^{id} is determined, for which the optimization problem is solved s times (by the number of partial criteria), each time with one (the next) criterion, as if the rest were not exist at all. The sequence of "single-criterion" solutions of the initial multicriteria problem gives the coordinates of an unattainable ideal vector $y^{id} = \left\{ y_k^{id} \right\}_{k=1}^{s}$.

After that, the criterion function $Y(y)$ is introduced as a measure of approximation to the ideal vector in the space of optimized criteria in the form of some non-negative function of the vector y^{id}-y, for example, in the form of a square of the Euclidean norm of this vector:

$$Y(y) = \left\| \frac{y^{id} - y}{y^{id}} \right\| = \sum_{k=1}^{s} \left[\frac{y_k^{id} - y_k}{y_k^{id}} \right]^2 .$$

The disadvantage of this method is the cumbersome procedure for determining the coordinates of an ideal vector. In addition, the possibility of violation of restrictions is not ruled out.

The choice of the scheme of compromises is carried out by the person making the decision (DM) and has a conceptual character.

Formalization

Depending on the availability and type of information on the preferences of DM, the approaches to solving multicriteria tasks can be different. If there is no such information at all, then sometimes we are limited to finding any solution vector x^* that ensures only the fulfillment of the constraints $A = \{A_k\}_{k=1}^{s}$ condition:

$$y^* \in M = \left\{ y \middle| 0 \le y_k\left(x^*\right) \le A_k, k \in [1, s] \right\}, \; x^* \in X.$$

(Here we have the structuring of the concept of the domain of constraints M).

The disadvantages are obvious – the solution obtained is often rough and, as a rule, not Pareto-optimal. Consequently, the capabilities of the system in this case are not fully used.

The method is recommended to be used to optimize very complex systems, when it is far from easy to carry out even such a simple reconciliation of conflicting criteria ("just to get into limitations"). A variation of this approach is the widely accepted technique, when for optimization of the set y_k, $k \in [1,s]$, the decision maker chooses only one criterion (for example, the first one), and the remaining criteria are reclassified into the category of constraints. Thus, the original multicriteria problem is artificially replaced by a one-criterion problem with constraints:

$$x^* = \underset{x \in X}{\operatorname{argmin}} \; y_1(x), 0 \le y_k(x) \le A_k, k \in [1, s].$$

A consequence of this approach is the solution in the form of a polar point of the Pareto region, i.e. frankly rude and subjective decision.

The scalar convolution approach with minimized criteria involves the use of the formula

$$x^* = \arg\min_{x \in X} Y[y(x)].$$

It is more reasonable in terms of formalization.

Analysis of the Scalar Convolution

The problem is that the form of the function $Y[y(x)]$ depends on the situation of the adoption of the multicriteria solution and is usually not known. Since the function $Y[y(x)]$ is difficult to obtain throughout the entire domain, we are often limited to an analysis of its behavior in the vicinity of that point in the arguments space that corresponds to the most typical situation. Since we are talking about *small* neighborhoods of the operating point, then, using the hypothesis of the smoothness of the criterion function, we replace it by a hyperplane tangent to the surface of equal values of $Y[y(x)]$ at the operating point. Then the approximating dependence $Y[\alpha, y(x)]$ takes the form of a linear scalar convolution

$$Y^o[a, y(x)] = \sum_{k=1}^{s} a_k^o y_k(x),$$

where α^o_k is the regression coefficient having the meaning of the partial derivative of the criterion function with respect to the k-th criterion, calculated at the base operating point. To calculate the coefficients α with the use of information from the DM, it is possible to solve the problem using the least squares method, but it is better to use the heuristic modeling technique described in [4].

Using the expression obtained, it must always be remembered that this is only a linear approximation of the scalar convolution of criterial functions, and in situations that differ from the base one, it can lead to significant distortions.

To obtain a criterion function over the entire domain, it is necessary to specify the form of the approximation dependence. As usual in the practice of approximation, success depends on how adequately the form of the given function reflects the physics of the phenomenon being studied. If you use information about the mechanisms of phenomena, then the model you specify is meaningful. In the absence of such information, the "black box" approach is used, and formal regression models of a general type (polynomial, power, etc.) are given for approximation. The quality of meaningful models is usually much better than formal ones.

Content Analysis of Utility Function

To improve the quality of the research, one should always involve a priori information about the physics of the phenomenon under investigation and, at every opportunity, move from formal models to meaningful ones. In our case, the subject of investigation is such a subtle substance as an imaginary utility function that arises in the mind of the decision-maker when solving a particular multicriteria problem. In addition, even if it does exist, then each DM has its own

utility function. Nevertheless, it is possible to obtain information for specifying the type of the meaningful model of a criterial function if one reveals and analyzes some general laws observed in the process of making multicriteria decisions by various decision-makers in different situations.

Comparison of partial criteria of a different physical nature is possible only in a normalized (dimensionless) space. We normalize the efficiency vector y by the constraint vector A and obtain a vector of relative partial criteria (the normalized efficiency vector)

$$y_0(x) = \{y_k(x)/A_k\}_{k=1}^{s} = \{y_{0k}(x)\}_{k=1}^{s}.$$

This operation is monotonic, and, in accordance with the well-known theorem of Hermeyer, any monotonic transformation does not change the results of the comparison. Therefore, we replace the model of the solution of the vector optimization problem with the original criterion functions by the model

$$x^* = \arg\min_{x \in X} Y[y_0(x)],\ y_{0k}(x) \in [0;1], k \in [1,s],$$

in which the practically used schemes of compromises have a physical meaning. The form of the function $Y[y_0(x)]$ depends on the chosen scheme of compromises.

The scheme of compromises determines in what sense the multicriteria solution obtained is better than other Pareto-optimal solutions. At present, the choice of the scheme of compromises is not determined by theory, but is carried out heuristically, on the basis of individual preferences and professional experience

of the developer, as well as information about the situation in which a multi-criteria decision is taken.

The main difficulty of the transition from the vector quality criterion to the scalar convolution is that the convolution should be a conglomeration of partial criteria, the significance (importance) of each of which in the overall assessment changes depending on the situation. In various situations, the rank of "the most important" can acquire different partial criteria. In other words, the scalar convolution of partial criteria must be an expression of a scheme of compromises that *depends on the situation*. When analyzing the possibilities of formalizing the choice of the scheme of compromises, let's put this thesis in the basis.

It is assumed that there are some invariants, rules that are usually common to all decision-makers, regardless of their individual inclinations, and which they equally adhere to in any given situation. The inevitable subjectivity of a decision maker has its limits [5]. In business decisions, a person must be rational in order to be able to convince others, explain the motives of his choice, the logic of his subjective model. Therefore, any preferences of decision-makers should be within the framework of a certain rational system. This makes possible formalization.

The concept of the situation, expressed by the deuce $S=<r,x>$ from the Cartesian product $R\times X$, is fundamental to the theory of vector optimization, since it, being objective, is the only support for attempts to formalize the choice of the compromise scheme. We introduce the concept of **tension of the situation** as a measure of the closeness of relative partial criteria to their limiting value (unit):

$$\rho_k(r,x)=1-y_{0k}(r,x),\ \rho_k\in[0;1],\ k\in[1,s].$$

This system is a structured characteristic of the concept of the situation $S=<r,x>$, $r \in R$, $x \in X$.

If a multicriteria solution is taken in a **stressful** situation, then it means that under given external conditions r, one or more partial criteria $y_{0k}(r,x), k \in [1,s]$, as a result of the solution x, may be in dangerous proximity to the limiting value ($\rho_k = 0$). And if one of them reaches the limit (or exceed it), then this event is not compensated by a possible low level of the remaining criteria – it is usually not allowed to violate any of the restrictions.

In this situation, it is necessary to prevent in every possible way the dangerous growth of the most unfavorable (i.e., closest to its limit) partial criterion, not taking very much into account the behavior of the others at this time. Therefore, in sufficiently stressful situations (for small values of ρ_k), the DM, if it admits the deterioration of the maximal (most important in the given conditions) partial criterion per a unit, then only compensating by a large number of units for improving the remaining criteria. And in a very tense situation (the first polar case: $\rho_k \approx 0$), the DM generally leaves only this one, the most unfavorable partial criterion, in view, without paying attention to the others.

Consequently, an adequate expression of the scheme of compromises in the case of a stressful situation is the minimax (Chebyshev) model

$$x^* = \arg\min_{x \in X} \max_{k \in [1,s]} y_{0k}(x). \qquad (1)$$

In less stressful situations, it is necessary to return to the simultaneous satisfaction of other criteria, taking into account the contradictory unity of all interests and goals of the system. In this case, the DM varies his estimate of the winnings according to one criteria and the losses on the other, depending on the situation. In intermediate cases, schemes of compromises are chosen, giving different degrees of partial equalization of partial criteria. With a decrease in the tension of the situation, preferences for individual criteria are aligned.

And, finally, in the second polar case ($\rho_k \approx 1$) the situation is so calm that the partial criteria are small and there is no threat of violation of the restrictions. DM here considers that the unit of deterioration of any of partial criteria is completely compensated by an equivalent unit of improvement of any of the others. This case corresponds to an economical scheme of compromises, which provides the minimum for the given conditions, the total losses by partial criteria. Such a scheme is expressed by the model of integral optimality

$$x^* = \arg\min_{x \in X} \sum_{k=1}^{s} y_{0k}(x). \qquad (2)$$

Analysis shows that schemes of compromises are grouped at two poles, reflecting different principles of optimality: 1) egalitarian – the principle of uniformity and 2) utilitarian – the principle of economy.

The application of the principle of uniformity expresses the aspiration uniformly, i.e. equally reduce the level of all relative criteria in the functioning of the management system. A very important realization of the principle of uniformity is the Chebyshev model (1) – the polar scheme of this group. This scheme makes it necessary to minimize the worst (greatest) of the relative criteria, reducing it to the level of the others, i.e. leveling all the partial criteria. The

disadvantages of egalitarian schemes of uniformity include their "economic inefficiency". Providing the closest to each other level of relative criteria is often achieved by significantly increasing their total level. In addition, sometimes even a small digression from the principle of uniformity can significantly improve one or more important criteria.

The principle of economy, which is based on the possibility of compensating for some deterioration in quality according to one criteria with a certain improvement for others, is devoid of these shortcomings. The polar scheme of this group is realized by the model of integral optimality (2). The utilitarian scheme provides the minimum total level of relative criteria. A common drawback of schemes of the principle of economy is the possibility of a sharp differentiation of the level of individual criteria.

The analysis reveals a pattern by which the decision maker varies from a model of integral optimality (2) in calm situations to a minimax model (1) in stressful situations. In intermediate cases, the decision maker chooses compromise schemes that give different degrees of satisfaction of individual criteria in accordance with his individual preferences, but in accordance with the given situation. If we take the conclusions from the analysis as a logical basis for formalizing the choice of the compromise scheme, then we can suggest various constructive concepts, one of which is the concept of a nonlinear scheme of compromises.

Nonlinear Scheme of Compromises

From the standpoint of the systemic approach, it is advisable to replace the problem of *choosing* the scheme of compromises with the equivalent problem of synthesizing a certain *single* scalar convolution of partial criteria, which in different situations would express different principles of optimality. Separate

models of compromise schemes are combined into a single holistic model, the structure of which adapts to the situation of a multi-criteria decision making. The requirements for the synthesized function $Y[y_0(x)]$:

- it should be smooth and differentiable;
- in tense situations, it should express the principle of minimax;
- in calm conditions – the principle of integral optimality;
- in intermediate cases it should lead to Pareto-optimal solutions, giving various measures of partial satisfaction of the criteria.

In other words, such a universal convolution should be an expression of a scheme of compromises that *adapts* to the situation. We can say that adaptation and the ability to adapt are the main substantive essence of the study of multi-criteria systems. For this it is necessary that the expression for the scalar convolution explicitly include the characteristics of the tension of the situation. We can consider several functions that satisfy the above requirements. The simplest of these is a scalar convolution

$$Y(\alpha, y_0) = \sum_{k=1}^{s} \alpha_k [1 - y_{0k}(x)]^{-1}; \alpha_k \geq 0, \sum_{k=1}^{s} \alpha_k = 1,$$

where α_k=const are the formal parameters defined on the simplex and having a double physical meaning. On the one hand, these are the coefficients that express the preferences of the decision-maker for certain criteria. On the other hand, these are the coefficients of regression of a meaningful regression model based on the concept of a nonlinear scheme of compromises.

Thus, the nonlinear scheme of compromises is considered as the basic one, to which corresponds the model of vector optimization, which explicitly depends on the characteristics of the tension of the situation:

$$x^* = \arg\min_{x \in X} \sum_{k=1}^{s} \alpha_k \left[1 - y_{0k}(x)\right]^{-1}. \qquad (3)$$

From this expression it is clear that if any of the relative partial criteria, for example, $y_{0i}(x)$, approaches closely to its limit (unit), i.e. the situation becomes strained, then the corresponding term $Y_i=\alpha_i/[1-y_{0i}(x)]$ in the minimized sum increases so much that the problem of minimizing the entire sum is reduced to minimizing only the given worst term, i.e., ultimately, the criterion $y_{0i}(x)$. This is equivalent to the action of the minimax model (1). If the relative partial criteria are far from unit, i.e. the situation is calm, then the model (3) acts equivalent to the model of integral optimality (2). In intermediate situations, different degrees of partial alignment of the criteria are obtained.

This means that the nonlinear compromise scheme has the property of continuous adaptation to a multi-criteria decision making situation. From this point of view, traditional schemes of compromises can be considered as a result of the "linearization" of a nonlinear scheme at various "work points" - situations. This, by the way, explains the name of the proposed *nonlinear* scheme of compromises, since in other respects it is no more "nonlinear" than other schemes considered in decision theory. We emphasize that the adaptation of the nonlinear scheme to the situation is carried out *continuously*, while the traditional choice of the compromise scheme is done discretely, which adds to the subjective errors also the errors associated with the quantization of the compromise schemes.

We have repeatedly stressed above that the choice of a compromise scheme is a person's prerogative, a reflection of his subjective utility function when solving a particular multicriteria task. Nevertheless, we managed to identify some regularities and, on this objective basis, construct a scalar convolution of

criteria, the form of which follows from meaningful ideas about the essence of the phenomenon under study. The phenomenon of individual preferences of the DM is formally represented by the presence of the vector α in the structure of the meaningful model (3).

The Pareto optimality questions of the nonlinear scheme of compromises and its axiomatics were investigated in [4,6].

Unification

Various assessments of the role of subjective factors in the solution of multicriteria problems are possible. Subjectivity is permissible and even desirable if such a task is solved in the interests of a particular person. Therefore, the mechanism of individual preferences is rather intensively applied in the practice of solving multicriteria problems.

However, subjectivity in their decision is permissible and desirable only as long as the result is intended for specific decision-makers or narrow collectives of people with similar preferences. If it is intended for general use, then it must be completely objective, unified.

When the result of solving a multicriteria problem is intended for wide use, it is unified and individual preferences are leveled by statistics. If there is no a priori information about the differentness of the criteria, then the principle of the insufficient foundation of Bernoulli-Laplace says that in this case we must accept all the weight coefficients in expression (3) *equal to each other*. It follows from the normalization on the simplex that $\alpha_k \equiv 1/s,\ \forall k \in [1,s]$. Then

$$Y(\alpha, y_0) = \frac{1}{s} \sum_{k=1}^{s} [1 - y_{0k}(x)]^{-1}.$$

Taking into account that multiplication by $1/s$ is a monotonic transformation, which, by the theorem of Hermeyer, does not change the results of the comparison, we pass to the unified (without weight coefficients) expression for the scalar convolution of the criteria

$$Y(y_0) = \sum_{k=1}^{s} [1 - y_{0k}(x)]^{-1}. \quad (4)$$

This formula is recommended to be applied in all cases when a multicriteria problem is solved not in the interests of one particular DM, but for wide use.

The unified scalar convolution by the nonlinear scheme has the form (4) or, in an equivalent form,

$$Y(y) = \sum_{k=1}^{s} A_k [A_k - y_k(x)]^{-1},$$

i.e., without preliminary normalization of partial criteria. The concept of a non-linear scheme of compromises corresponds to the principle "away from restrictions".

For the criteria being maximized, the unified scalar convolution has the form

$$Y(y) = \sum_{k=1}^{s} B_k [y_k(x) - B_k]^{-1},$$

where $B_k, k \in [1, s]$ is the minimum permissible values of the criteria to be maximized.

The Dual Method

If a multicriteria problem is solved in the interests of a particular decision maker, it is recommended first to obtain a unified (basic) solution and present it to the person. And only if this solution does not satisfy him and correction is required, it is necessary to proceed to the determination of weight coefficients reflecting his individual preferences. It is important that the search process does not start from an arbitrary point in the criteria space, but from a unified, basic solution.

The practice of solving multicriteria problems shows that the assumption that there is a ready and stable (at least implicitly) utility function of the decision maker is not always fair. Solving a multi-criteria task, the decision maker compares sets of specific criteria values with different alternatives, makes trial steps, makes mistakes and interprets the relationship between his needs and the possibilities of meeting them with a given object in a given situation.

With contradictory criteria, this ratio is by its nature a compromise, however, a decision maker does not have a consciously a priori scheme of compromises, or so far it is only in its infancy. Usually, the idea of a compromise scheme that is necessary to solve a problem arises and is gradually improved only as a result of attempts by the decision maker to improve a multi-criteria solution in a

series of test steps. It is clear that the presence of interactive computer technology is implied. "In kind" such a procedure is usually impossible.

Thus, simultaneously and interdependent, on the one hand, a person adapts to the multicriteria problem being solved, carrying out the structuring of preferences and improving his understanding of the utility function, and on the other, consistently finds a series of solutions optimal in the sense of the current utility function. The mutually conditioned processes of adaptation of the decision maker to the task and finding the best result are of a dual nature and are, in principle, part of the methodology of the human-machine solution of multicriteria problems.

As noted, in the initial stage of the decision process, the DM practically lacks not only an analytic description of the utility function, but also a ready a priori idea of it. Therefore, the interactive procedure should be organized as dual, and the search optimization method should allow dialog programming in ordinal scales and use minimal information about the utility function. This method, based on the comparison of preferences with specially calculated alternatives, is an ordinal analogue of the simplex-planning method [4,6].

An important factor contributing to the effectiveness of the method is that the starting point of the search is chosen not as an arbitrary point in the Pareto set, but as an axiomatically grounded basic solution that should only be adjusted in accordance with the informal preferences of a particular decision maker. The process of adjustment provides mutual adaptation: a person adapts to this particular multicriteria task, and the model of a non-linear scheme of compromises becomes a reflection of the individual preferences of the person.

The fundamental difference between convolution in a nonlinear scheme and other known scalar convolutions is the organic connection with the situation of a multi-criteria decision. In fact, the proposed convolution is a non-linear

regression function (linear in parameters), chosen for physical reasons and therefore effective. The coefficients α in the expression for the nonlinear scalar convolution have the meaning of the parameters of the nonlinear meaningful regression function, therefore, when found, they do not change from situation to situation, as in the case of linear and other known convolutions that do not adapt to the situation.

The problem of determining the coefficients α in a dual procedure can be considered as the problem of synthesizing a decision rule, which, when applied formally, adequately reflects the logic of a particular decision maker in any possible situation. Such a problem arises, for example, when a multi-criteria system operates in the mode of the operator's advisor in the conditions of time deficit. Here, it is desirable that the system in any situation quickly made the same decision as this operator, if he had the opportunity to calmly think. Similar problems have to be solved in the development of a decisive system for an intelligent robot that functions in changing and uncertain dynamic environments, if you want it to act in the same way a person who trained it would act in its place, etc.

Conclusion

As a result of a systemic approach, a multi-criteria optimization model is obtained, which allows an object to achieve all of its goals in the whole range of possible situations. A systemic approach to the problem of multi-criteria optimization allowed us to combine the models of individual compromise schemes into a single holistic structure, adapting to the situation of multi-criteria decision making. The advantage of the concept of a non-linear compromise scheme is the possibility of making a multi-criteria decision formally, without direct human participation. At the same time, on a single ideological basis, both tasks that are important for general use and those which main content essence is the satisfaction of individual preferences of decision makers are solved. The apparatus of the non-linear compromise scheme, developed as a formalized

tool for the study of control systems with conflicting criteria, allows us to practically solve multi-criteria tasks of a wide class.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

1. Antonov A.V. System analysis: a textbook for universities. - M .: Higher School, 2004. - 454 p.
2. Peregudov F.I., Tarasenko F.P. Introduction to system analysis. - M .: Radio and communication, 1989. - 320 p.
3. Fishburn P. Theory of utility for decision-making. - M.: Science, 1978. - 352 p.
4. Voronin A.N., Ziatdinov Yu.K., Kuklinsky M.V. Multi-criteria solutions: Models and methods. - K.: NAU, 2010.– 348 p.
5. Larichev O.I. Science and the art of decision making. - M: Science, 1979. - 200 p.
6. Albert Voronin. Multi-Criteria Decision Making for the Management of Complex Systems. – USA: IGI Global, 2017. – 201 p.

Authors' Information

Voronin Albert Nikolaevich *– professor, Dr.Sci (Eng), professor of the department of National aviation university of Ukraine, Lubomyr Husar Avenue 1, 03680, Kyiv, Ukraine; e-mail: alnv@ukr.net*

Major Fields of Scientific Research: Multi-Criteria Decision Making; Man-Machine Complex Systems

Ziatdinov Yuriy Kashafovich – *professor, Dr.Sci (Eng), dean of the aerospace faculty of National aviation university of Ukraine, Lubomyr Husar Avenue 1, 03680, Kyiv, Ukraine; e-mail: oberst5555@gmail.com*

MULTISENSOR PROTOTYPE FOR BEVERAGE QUALITY CONTROL: PRINCIPLE SCHEME AND TEST RESULTS

Volodymyr Romanov, Igor Galelyuka, Oleksandr Voronenko, Oleksandra Kovyrova, Sergei Dzyadevych, Lyudmyla Shkotova

***Abstract**: Nowadays there is no general theory of wireless sensor networks, and existing mathematical models are fragmented. Therefore, together with the development of theoretical foundations and testing them on appropriate mathematical models, it is necessary to work out design solutions for multisensors and sensor networks on physical models or prototypes. It is envisaged that the multisensor will be delivered to the market with a set of interchangeable sensor modules with ability of quick replace them when moving from one series of measurements to another in order to speed up the process of controlling beverage quality at all stages of production. The results of testing circuit solutions and test methods based on the prototype multisensor for real-time monitoring of the quality of beverages are presented in this publication.*

***Keywords**: wireless smart multisensor networks, test of beverage quality*

***ITHEA Keywords**: J.3 Life and Medical Sciences- Biology and Genetics.*

Introduction

"Smart" multisensors and biosensor systems based on modern information and communication technologies make it possible to qualitatively improve the parameters of systems for testing biologically active, chemical and toxic substances, biological or biophysical objects, and improve the parameters control, processing and analysis of data in the food industry, specifically in beverage production, digital agriculture, environmental monitoring and other areas of human activity. It should be noted that sensor technologies are one of

the key world technologies, the development of which is exponential, that is, the parameters of these technologies improve by tens or even hundreds of percent per year. In addition, if it is possible to use such technologies in conjunction with, for example, IT and microelectronics then there is a programmed technological explosion. According to IoT Analytics [Smart, 2019], in 2019, 14% of all meters are now smart meters. The market of “smart” meters increases every year and this increasing is predicted for many years further. This market includes smart sensors and smart devices that can operate individually or as part of wireless or wired sensor networks, for measuring electrical parameters as well as parameters of the liquid and gas environment. In the next two years, this market is expected to increase by another 1 billion smart sensors, including bio- and multisensors. The main idea of the article is to show the results of circuit design and testing prototype of wireless "smart" multisensors based on amperometric enzyme biosensors to control the quality of beverages including during their production.

Work objectives

Work objectives are design and prototyping the smart biological sensor with wireless data communication unit for quality control of beverages.

Smart biosensor design

The world's leading companies of electronic components market have been working successfully over the last few years to create new microcontrollers designed to work with electrochemical modules including bio- and multisensors. The leader among such companies is the Analog Devices Company, which at the beginning of 2019 put into production a new chip of the microcontroller ADuCM355 [Analog, 2020]. The block diagram of the microcontroller is shown in Figure 1.

This microcontroller is designed to work with electrochemical and biosensors. It is based on the ARM® Cortex®-M3 processor core and can operate in current,

voltage and resistance measurement modes. The microcontroller contains a 16-bit ADC, filter, amplifier with programmable gain, trans-impedance amplifier with programmable gain to connect sensors of different types, three DACs with outputs on voltage, direct access controller to independent interfaces, UART port and I2C interface.

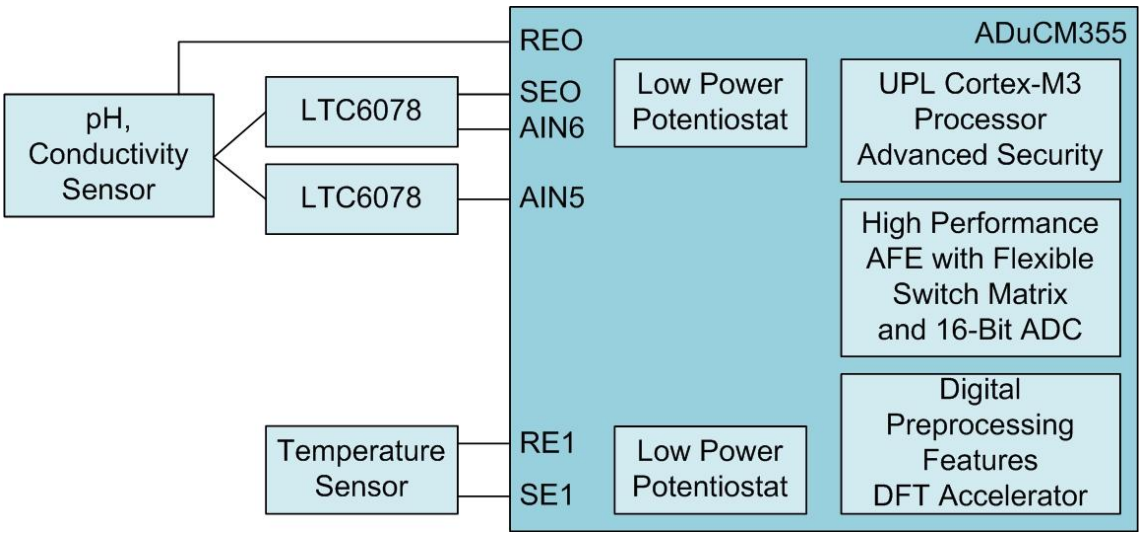

Figure 1. Block diagram of the ADuCM355 microcontroller

Most electrochemical and biosensors can be directly connected to the microcontroller. Additional LTC6078 type amplifiers can be used to increase the sensitivity of the measuring channel. This is a dual-rail input-output CMOS low-noise amplifier with low power consumption. The amplifier has large input impedance that allows it to be successfully matched to the high output impedance of the biosensor to provide the required measurement accuracy. When working with electrochemical or biosensors, it is usually necessary to control the ambient temperature in order to compensate the sensor temperature error. Such microcontroller also allows the measurement of the impedance of the sensors in the range from 100 Ohms to 10 MOhms. The large dynamic range is especially important for determining the electrical conductivity, which allows measuring different concentrations of the test solutions. It should be

noted that multisensory prototype was created on the base of the previous microcontroller ADuCM350, which is fully compatible with the new ADuCM355.

Multisensor functional diagram

The multisensor functional diagram with the main blocks is shown in Figure 2.

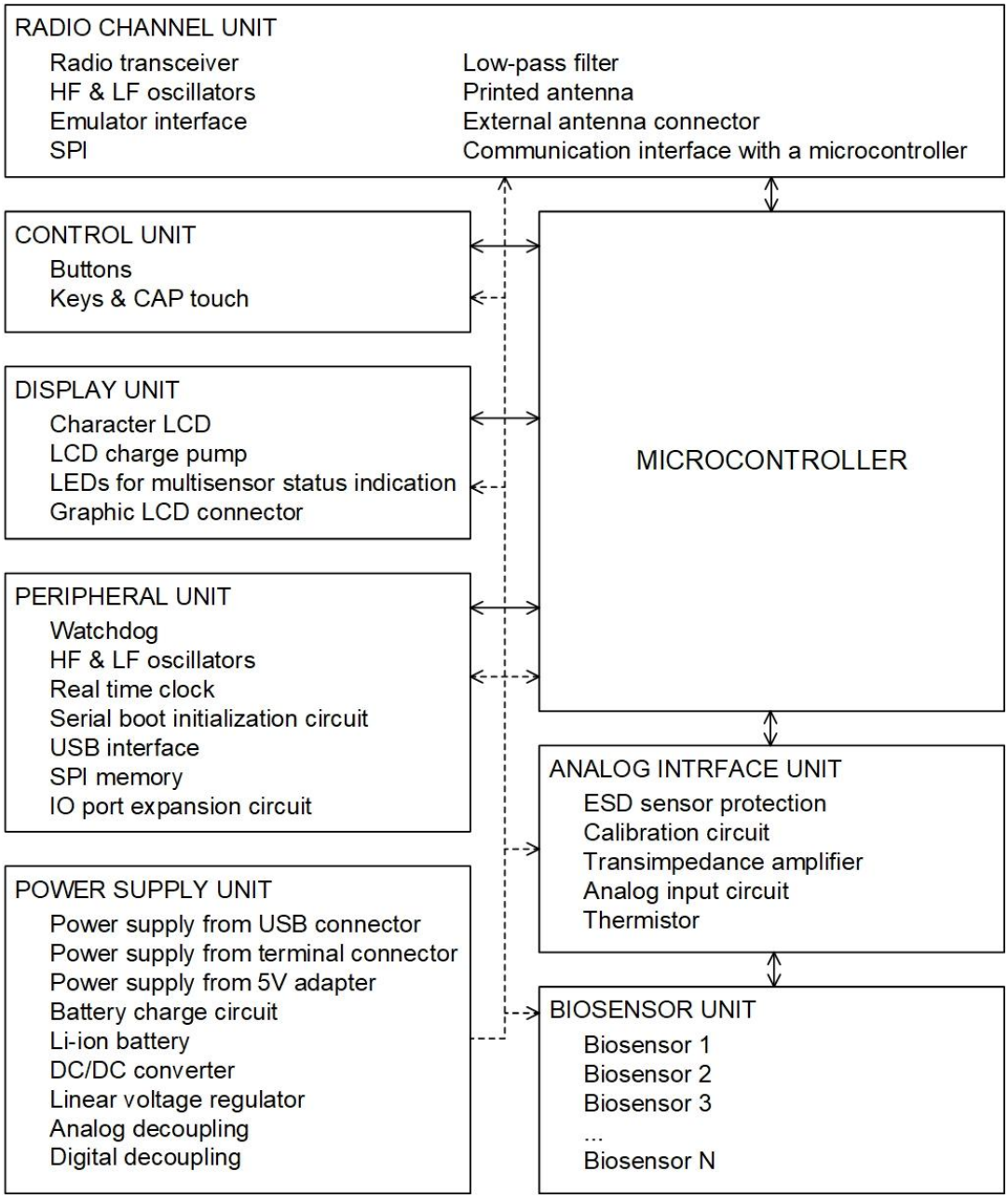

Figure 2. Multisensor functional diagram

The multisensor functional diagram in addition to the microcontroller based on the ADuCM350, contains the following units:

- The control unit that contains the keyboard or buttons, which allows to check the main functions of the multisensor both from the biosensors and the simulator offline;
- The display unit used when setting up the microcontroller or when using the multisensor as a standalone device;
- The peripheral unit that contains the necessary interfaces for the data exchange with internal and external means;
- The power supply unit that supplies all multisensor blocks with a given voltage level;
- ADuCM350 microcontroller unit;
- The analog interface unit designed to communicate with individual biosensors;
- The biosensor units;
- The radio channel unit designed to organize a wireless multisensor network.

Acting multisensor prototype

Acting multisensor prototype, which, in addition to Figure 2, includes a programmer block, is shown in Figure 3.

The created multisensor prototype was tested when working with different types of biosensors and solutions [Shkotova et al, 2016] to get recommendations for improving the scheme. The multisensor wireless channel was tested separately. In Figure 4, there is a graph of experimental results when measuring glucose concentration in a buffer solution.

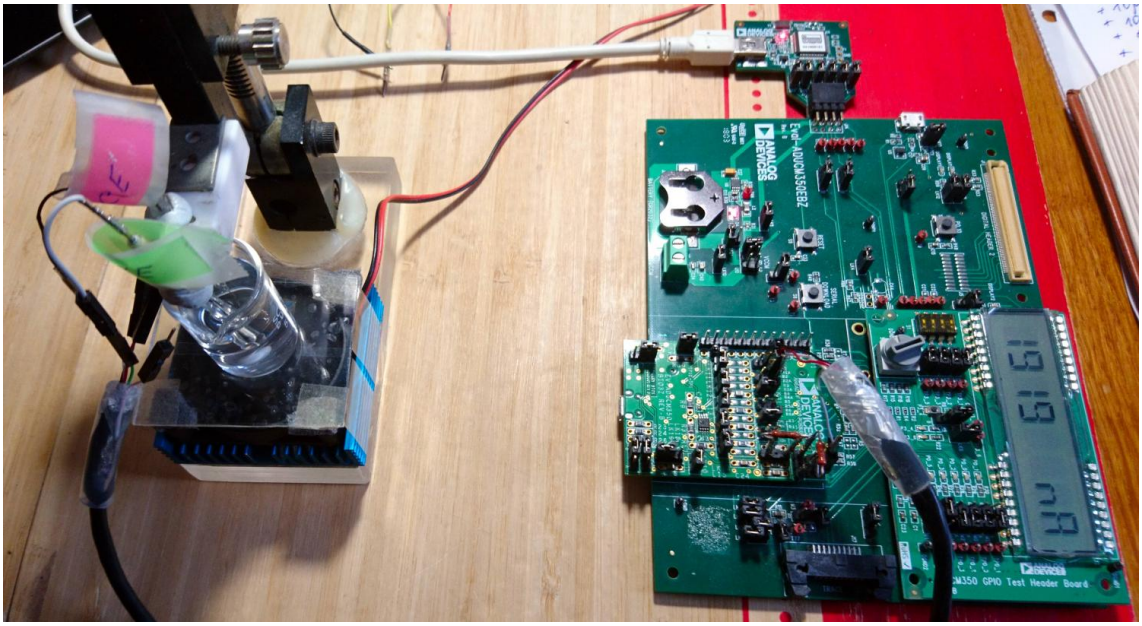

Figure 3. Acting multisensor prototype

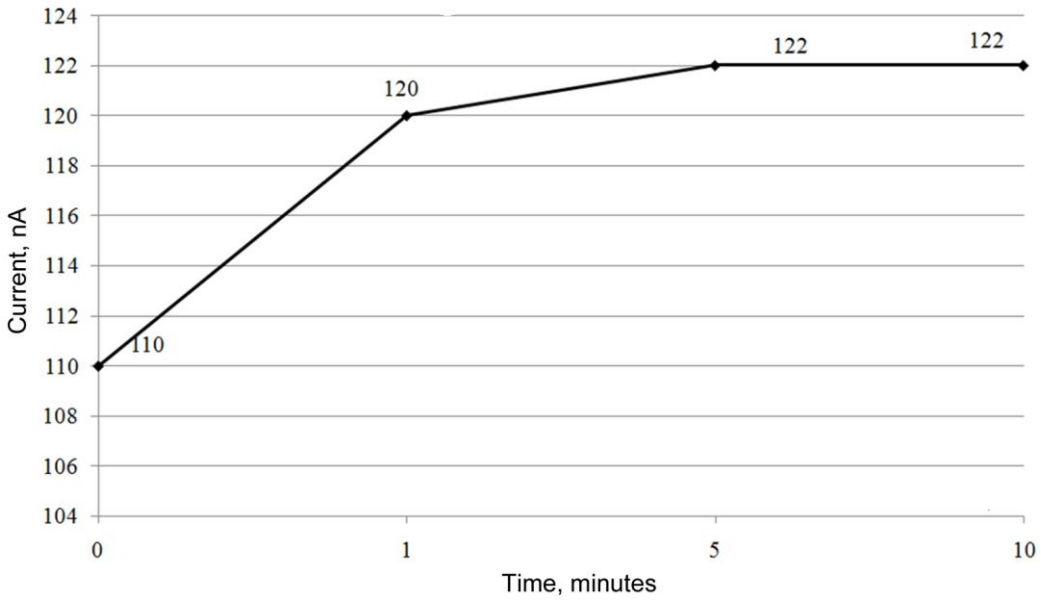

Figure 4. The results of test of measuring the 0.25 mmol of glucose in the 5 ml of buffer solution

Biosensor network

Previously, all wireless nodes of the prototype network in our development were built on the base of the wireless microcontroller JN5168 [Romanov et al, 2019]. By this structure, each node in the network, including the multisensors and the coordinator, contained a 32-bit, 32 MHz, RISC-processor and IEEE802.15.4 wireless transceiver. The network used ZigBee Pro stack as a wireless protocol for networking. The main unit for managing such network is the network coordinator. In addition, the coordinator supports the following functions: collecting, processing, visualizing and transmitting data to a workstation, Internet or cloud environment.

However, as identified by analyzing different uses of the network to evaluate the quality of different beverages, the use of the ZigBee industrial protocol limits the ability of network multisensors. This protocol requires the use of ZigBee/USB converters when using portable computers to manage the network. There are no USB ports in mobile tablets and in modern mobile phones, but at the same time these devices include standard Wi-Fi or Bluetooth wireless adapters, so the use of these ports allows replacing the coordinator in a number of applications with a tablet or mobile phone. Therefore, in the new wireless network for quality control of different beverages, it will be used not only ZigBee but also Bluetooth 5 protocol. The Bluetooth 5 protocol is focused on low power devices, which is important for the network with a battery power source. The main advantages of Bluetooth 5 (and later versions) are following: data exchange is supported at distances of up to 150 m (in open environment), which practically coincides with the capabilities of ZigBee. In industrial premises, the Bluetooth 5 range is 30-35 m. For this purpose, the Nordic Semiconductor nRF52840 (Figure 5) microcontroller was used to support the Bluetooth 5 protocol.

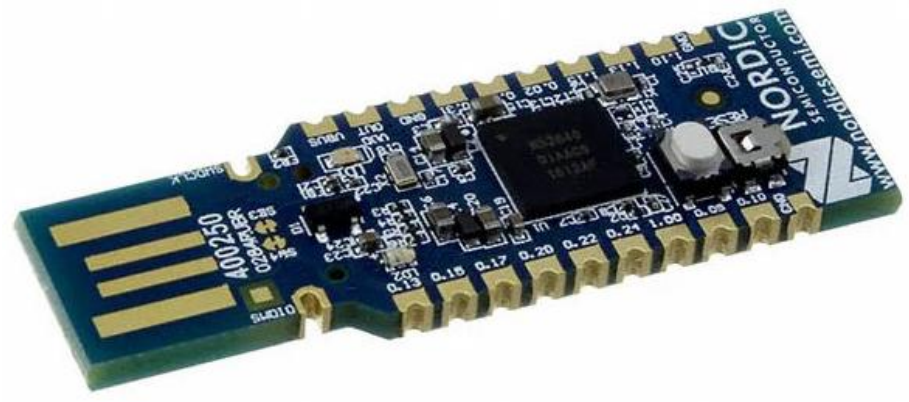

Figure 5. Wireless microcontroller nRF52840

The microcontroller nRF52840 is a multi-protocol device that supports the protocols Bluetooth 5, Thread, ZigBee, 802.15.4 and others. It is based on a 32-bit floating point ARM Cortex-M4 processor and has a clock speed of 64 MHz. Protocol switching is performed automatically without software restart. The additional use of Bluetooth has greatly expanded the capabilities of the developed multisensor. Thus, wireless multisensors can work as part of a network using ZigBee or standalone using Bluetooth protocol. In standalone mode, wireless multisensors can operate under control of mobile phone or tablet. The appearance of coordinator that supports a two-protocol sensor network is shown in Figure 6.

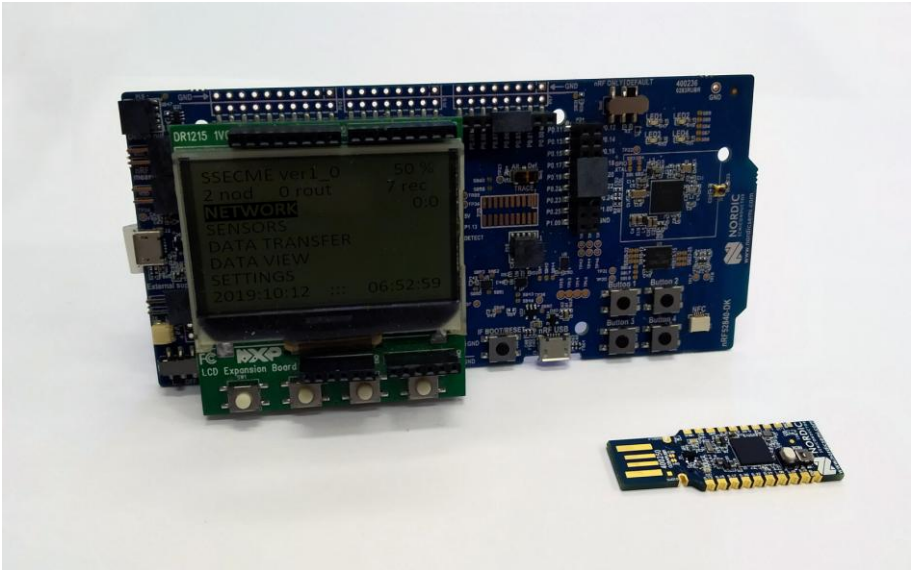

Figure 6. New coordinator (without housing) for wireless multisensor network

Conclusion

Features of circuit design solutions of multisensor module for quality control of beverage are considered. Examples of testing the concentration of glucose in solution using a prototype multisensor module are given. Features of the new network coordinator are disclosed.

Acknowledgement

The work is supported by the National Academy of Sciences of Ukraine in the frame of Scientific and Technical Program "Smart sensor devices of a new generation based on modern materials and technologies".

Bibliography

[Smart, 2019] Smart Meter Market 2019: Global penetration reached 14%, 2019, November. https://iot-analytics.com/smart-meter-market-2019-global-penetration-reached-14-percent (last accessed: February 2020).

[Analog, 2020] Precision Analog Microcontroller with Chemical Sensor Interface ADuCM355. https://www.analog.com/en/products/aducm355.html (last accessed: February 2020).

[Romanov et al, 2019] Romanov V., Galelyuka I., Voronenko O., Kovyrova O., Dzyadevych S., Shkotova L. Wireless smart multisensor networks for winemaking process control. International journal "Information theories and applications", Volume 26, Number 2. Sofia, Bulgaria. 2019. pp. 165–177. ISSN 1310-0513 (printed).

[Shkotova et al, 2016] Shkotova L., Pechniakova N., Kukla O., Dzyadevych S. Thin-film amperometric multibiosensor for simultaneous determination of lactate and glucose in wine. Food Chemistry. 2016, 197. pp. 972-978.

Authors' Information

Volodymyr Romanov *– Head of department of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of engineering sciences, professor; Prospect Akademika Glushkova 40, Kyiv, 03187, Ukraine;*

e-mail: VRomanov@i.ua;

website: http://www.dasd.com.ua

Igor Galelyuka *– Leading research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of engineering sciences; Prospect Akademika Glushkova 40, Kyiv, 03187, Ukraine;*

e-mail: galelyuka@gmail.com;

website: http://www.dasd.com.ua

Oleksandr Voronenko *– Research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kyiv, 03187, Ukraine;*

e-mail: alexander.voronenko@dci.kiev.ua;

website: http://www.dasd.com.ua

Kovyrova Oleksandra *– Research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kyiv, 03187, Ukraine;*

e-mail: kovyrova.oleksandra@gmail.com;

website: http://www.dasd.com.ua

Dzyadevych Sergei *– Deputy Director of Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine; Zabolotnogo Str., 150, Kyiv, 03143, Ukraine,*

website: http://www.imbg.org.ua/en/

Shkotova Lyudmyla *– Senior scientist of Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine; Zabolotnogo Str., 150, Kyiv, 03143, Ukraine,*

website: http://www.imbg.org.ua/en/

A COMPARATIVE EXAMINATION OF CONVOLUTIONAL AUTOENCODER AND DENSENET APPLICATIONS FOR BREAST CANCER CLASSIFICATION

Naderan Maryam, Yuri Zaychenko

Abstract: *Breast cancer is one of the most widespread and dangerous cancers among women. It is a disease that requires quick and accurate diagnoses. For this task, convolutional neural networks (CNNs) represent a huge breakthrough in image recognition. Many CNNs however, require large datasets for training which is not always available to researchers. In this paper, the effects of using a small dataset will be compared between our proposed convolutional autoencoder (CA) and DenseNet. It is shown that when using a small dataset there is considerable overfitting when using DenseNet, whereas overfitting does not occur with the proposed CA. Moreover, the training time of the CA was faster than DenseNet, and sensitivity (recall) of the proposed model was 90%.*

Key words*: Breast cancer detection, convolutional autoencoder, DenseNet, image classification*

1. Introduction

Breast cancer is a serious public health risk, affecting a large number of women every year**.** According to the last studies, one in every eight women in the United States will develop breast cancer in her lifetime. In the last year, it is estimated there were 268,600 new cases of invasive breast cancer and 62,930 new cases of non-invasive breast cancer diagnosed in women in the U.S [1]. Breast cancer is one of the largest public health threats requiring early detection to save lives. When it comes to cancer detection, a false negative might leave a patient with a lack of treatment which can have dire consequences for the patient. A false positive will just lead to more testing and analysis, which will

eventually lead to the discovery of the false positive. For this reason, recall (sensitivity) is more important than accuracy.

The aim of this study is to develop deep learning methods that could improve the sensitivity and training time of diagnosing breast cancer. In order to reduce the training time of the model, the model should be simplified. However, if the convolutional DenseNet will be used, because the model is more complex, it requires more data to train the model. As a result, the time of training is increased. Alternatively, using a convolutional autoencoder, the number of convolutional layers will be chosen in a way that will simplify the model and reduce the chance of overfitting. Using a small sample size with the convolutional autoencoder, produced a better result than the DenseNet model while reducing the training time.

2. Review of previous works

Authors in [2] proposed a new method they call "end-to-end" method and used the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM). The authors compared two convolutional neural networks, VGG and Resnet50, with their own method. The accuracy achieved using their proposed model was 84% and 97% for cancer and no-cancer respectively and the sensitivity of their model was 86.1%.

In [3] two modified CNNs were proposed where the average accuracy was 85%. During the experiment, 2420 mammography scans were used for training their proposed models. However, the model was very complex.

Authors in [4] proposed a modified model, where a CNN was used for feature extraction, but the gradient boosted tree model was used for classification. The experiment used 1804 mammography scans, from the Digital Database for Screening Mammography (DDSM). This modified model achieved 85% accuracy at detecting images with masses, with a sensitivity of 85%. The weakness of both [3,4] is that the training time of the model is increased. In this paper, a modified model with a decreased number of parameters is proposed. This model reduces the training time, while maintaining or improving sensitivity.

3. Dataset

The BreakHist dataset was used for this experiment. The dataset includes two classes of tumors Malignant and Benign, which are further organized by tumor type. The dataset is also separated into four magnification zooms 40X, 100X, 200X and 400X. Fig. 1 illustrates some input images that were used for training the model. Figures 1.a - 2.d belong to the benign category and figures 1.e - 1.h belong to the malignant category.

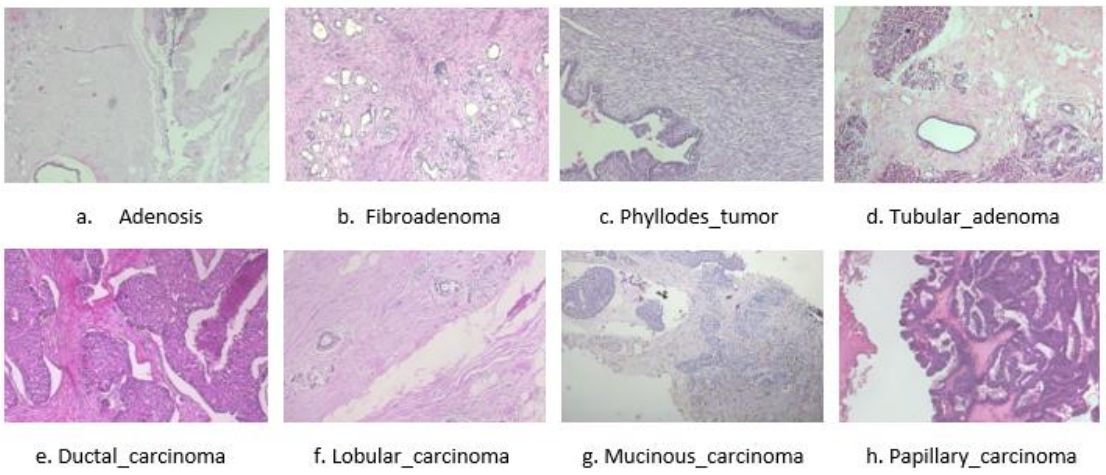

Fig. 1 Sample of dataset.

4. Experimental Investigations and Analysis

In this paper, all experiments were developed using Jupyter Labs, TensorFlow 2 and Python 3. The programs were implemented on a virtual machine with eight Intel CPUs.

Two convolutional networks DenseNet and Autoencoder were proposed in this work. The DenseNet model was both trained from scratch (FS) and fine-tuned (FT). There are some advantages of using the convolutional autoencoder:

1) It is challenging to get access to large datasets with labeled scans. However, since the autoencoder is an unsupervised model it does not require labeled

data, making it a viable solution for image recognition challenges involving a lack of labeled data, or small datasets.

2) The model is simple. Respectively, it has less parameters and as a result, the time of computation and training is drastically reduced.

The DenseNet (FS) model used 24 filters and 54 convolutional blocks. Each block included convolutional, activation, maxpooling and normalization layers. Therefore, the model is very complex and requires a large amount of data for training. Whereas, the convolutional autoencoder is simple with eight convolutional, eight batch normalization and two max-pooling layers. This simple architecture allows the CA to be trained with less data. Figure 2 illustrates the accuracy for the training and validation sets using DenseNet.

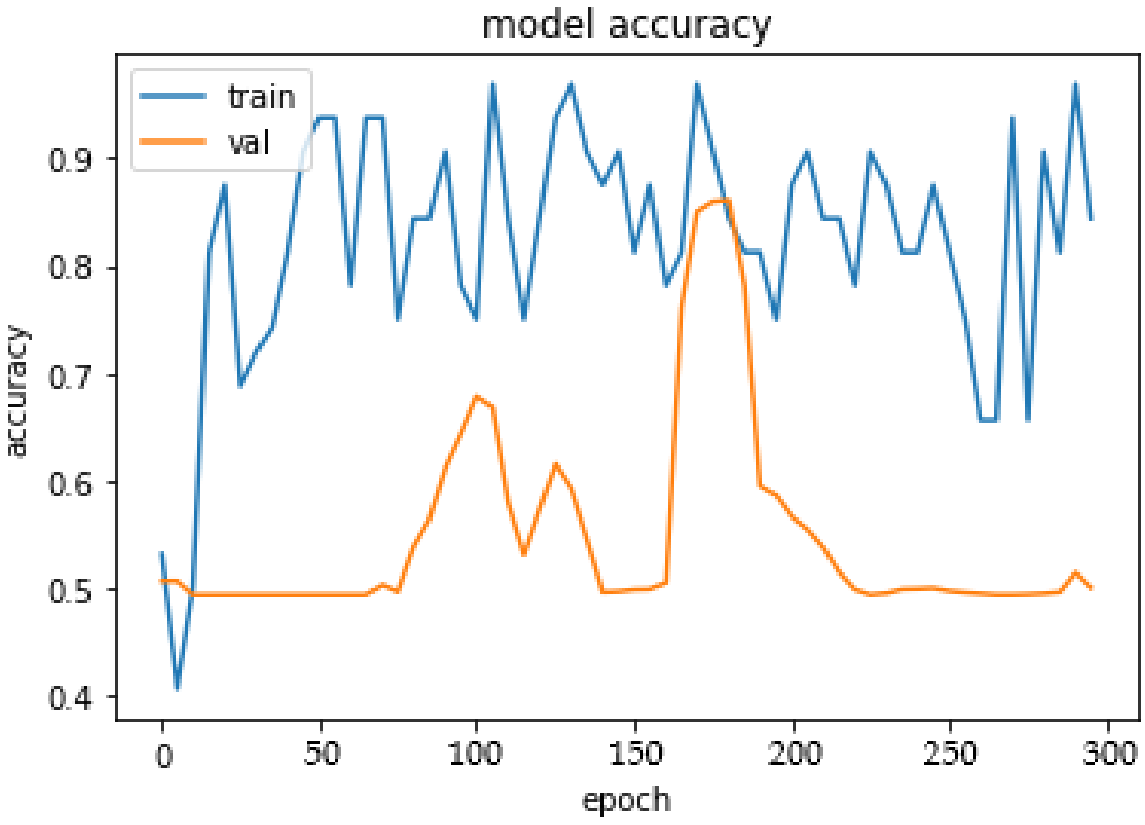

Fig. 2 Accuracy of training and validation using DenseNet

According to figure 2, it can be concluded that the model is overfitting since the accuracy for training data is significantly better than the accuracy for validation data. Using the same dataset, the modified convolutional autoencoder, showed better results than the DenseNet FS model. Figure 3 shows the loss function for the training and the validation set while using the autoencoder model.

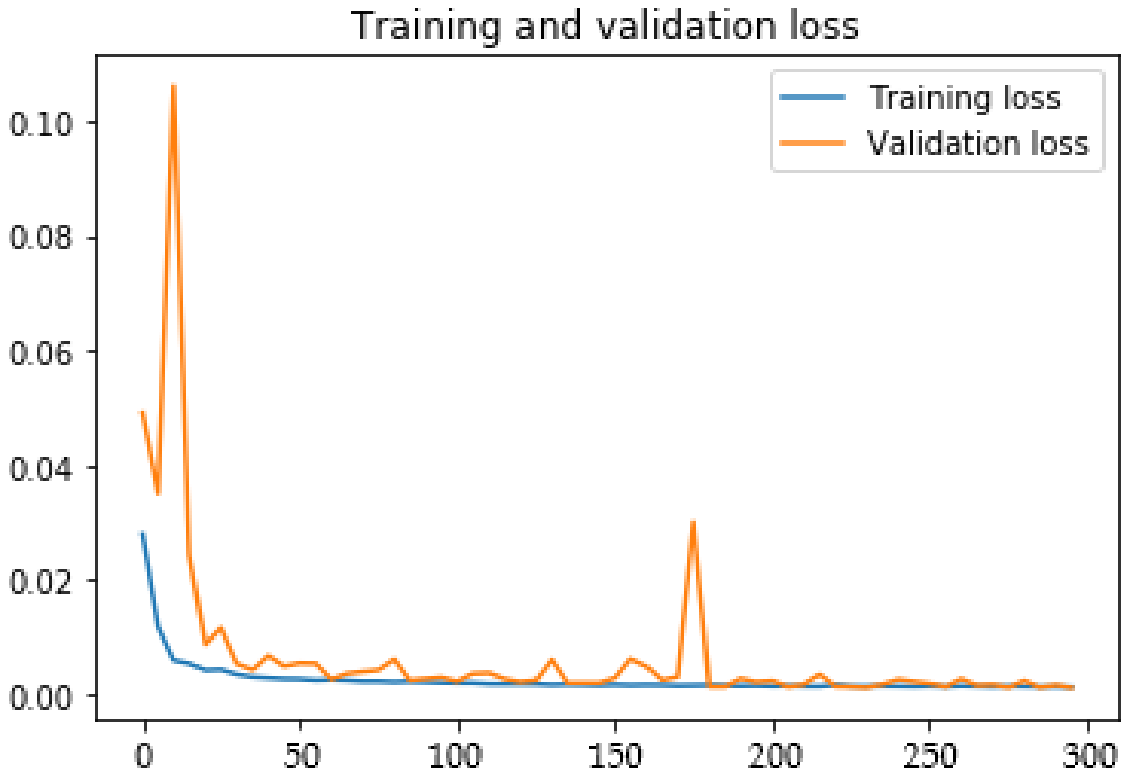

Fig. 3 Loss function of training and validation set using convolutional autoencoder

The loss function presents a measure of mistakes that were made by the network in predicting the output. Figure 3 also shows that after epoch 200, the value of the error function for training and validation data does not change and the loss reaches its minimum value. As a result, the optimal number of epochs in the current task is 200.

The DenseNet model was both trained from scratch and fine-tuned. To fine-tune the model, the output layer (Softmax layer) of pre-trained model was replaced with a new layer recognizing two classes, cancer and no cancer. The rest of the

pretrained layers were frozen. The sensitivity of this model as well as the proposed convolutional autoencoder were 90%, however the autoencoder required less data and training time.

Table 1 illustrates the results of the CNNs that were used in this paper.

Table 1 – Comparison of CNNs for Breast Cancer Detection

	Precision	Sensitivity	F1-Score	Accuracy
Convolutional Autoencoder	90.40%	90%	89.50%	90%
DenseNet FT	91%	90%	90.50%	95%
DenseNet FS	70%	65%	62.66%	67.50%

According to table 1, even though the sensitivity of the proposed CA and DenseNet FT are the same, training time in DenseNet FT was longer than CA. The training time of the DenseNet FT was about 3 hours, whereas in CA it was ~1.5 hours.

Conclusion

It is crucial to detect breast cancer at early stages, because if it is left undiscovered patients are at risk of more severe levels of cancer. For this reason, sensitivity is an important measurement in analyzing various CNNs. In our experiments we have shown that a high level of sensitivity (90%) is achieved with the convolutional autoencoder, while seeing other advantages as well. The convolutional autoencoder has less parameters than DenseNet, therefore, the model is less complex and prevents overfitting when using a small dataset. As a result, the training time of the model is dramatically decreased. In contrast, DenseNet requires a big amount of data to train the model and the training time increases considerably.

Bibliography

1 National Breast Cancer Foundation INC "Breast Cancer Facts" www.nationalbreastcancer.org/breast-cancer-facts.

2 Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, Weiva Sieh. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific Reports. (2019) 9:12495.

3 M. G. Ertosun and D. L. Rubin, "Probabilistic visual search for masses within mammography images using deep learning," *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Washington, DC, 2015, pp. 1310-1315.

4 Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. Med Phys 2017; 44: 1017–27. doi: https://doi.org/10.1002/mp.12110.

5 Zeshan Hussain, Francisco Gimenez, Darvin Yi, Daniel Rubin AMIA Annu Symp Proc. 2017; 2017: 979–984. Published online 2018 Apr 16. PMCID: PMC5977656.

Authors' Information

Naderan Maryam *- PhD student. Institute for applied system analysis, NTUU "KPI", 03056, Ukraine, Kyiv, Peremogi pr. 37, Corpus 35*

Yuri Zaychenko *– Professor, doctor of technical sciences,*

Institute for applied system analysis, NTUU "KPI", 03056, Ukraine, Kyiv, Peremogi pr. 37, Corpus 35; e-mail: baskervil@voliacable.com, zaychenkoyuri@ukr.net

Major Fields of Scientific Research: Information systems, Fuzzy logic, Decision making theory

TABLE OF CONTENTS